behavioral sciences

Special Issue Reprint

Educational Applications of Cognitive Psychology

Edited by
Jeri L. Little and Elizabeth Ligon Bjork

mdpi.com/journal/behavsci

MDPI

Educational Applications of Cognitive Psychology

Educational Applications of Cognitive Psychology

Guest Editors

Jeri L. Little
Elizabeth Ligon Bjork

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editors

Jeri L. Little
Department of Psychology
California State University
Hayward
USA

Elizabeth Ligon Bjork
Department of Psychology
University of California
Los Angeles
USA

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Behavioral Sciences* (ISSN 2076-328X), freely accessible at: https://www.mdpi.com/journal/behavsci/special_issues/U61884L3L8.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-8213-7 (Hbk)
ISBN 978-3-7258-8214-4 (PDF)
https://doi.org/10.3390/books978-3-7258-8214-4

© 2026 by the authors. Articles in this reprint are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The reprint as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

About the Editors . . . . . **ix**

Jeri L. Little and Elizabeth Ligon Bjork
A Special Issue Introduction: Educational Applications of Cognitive Psychology
Reprinted from: *Behav. Sci.* **2026**, *16*, 713, https://doi.org/10.3390/bs16050713 . . . . . **1**

Nate Kornell
Does Memory Accessibility Affect How Much We Learn from Studying?
Reprinted from: *Behav. Sci.* **2025**, *15*, 760, https://doi.org/10.3390/bs15060760 . . . . . **23**

Nate Kornell and Robert A. Bjork
Why Empirical Forgetting Curves Deviate from Actual Forgetting Rates: A Distribution Model of Forgetting
Reprinted from: *Behav. Sci.* **2025**, *15*, 924, https://doi.org/10.3390/bs15070924 . . . . . **35**

Alice F. Healy, Vivian I. Schneider and James A. Kole
Exploring Whether Making Second-Language Vocabulary Learning Difficult Enhances Retention and Transfer
Reprinted from: *Behav. Sci.* **2025**, *15*, 692, https://doi.org/10.3390/bs15050692 . . . . . **50**

Michael J. Serra, Althea N. Kaminske, Cynthia Nebel and Kristen M. Coppola
The Use of Retrieval Practice in the Health Professions: A State-of-the-Art Review
Reprinted from: *Behav. Sci.* **2025**, *15*, 974, https://doi.org/10.3390/bs15070974 . . . . . **69**

Spenser A. Clark, Michelle L. Rivers and Acacia L. Overono
The Role of Answer Justification in Multiple-Choice Testing: Effects on Performance and Metacognitive Accuracy
Reprinted from: *Behav. Sci.* **2025**, *15*, 477, https://doi.org/10.3390/bs15040477 . . . . . **98**

Amber E. Witherby, Paige E. Northern and Sarah K. Tauber
Does Covert Retrieval Benefit Adolescents' Learning in 8th Grade Science Classes?
Reprinted from: *Behav. Sci.* **2025**, *15*, 843, https://doi.org/10.3390/bs15070843 . . . . . **116**

Kelsey K. James and Benjamin C. Storm
Sorry, Am I Intruding? Comparing Performance and Intrusion Rates for Pretested and Posttested Information
Reprinted from: *Behav. Sci.* **2025**, *15*, 1060, https://doi.org/10.3390/bs15081060 . . . . . **137**

Zohara Assadipour, Dahwi Ahn and Jason C. K. Chan
Interpolated Retrieval of Relevant Material, Not Irrelevant Material, Enhances New Learning of a Video Lecture In-Person and Online
Reprinted from: *Behav. Sci.* **2025**, *15*, 668, https://doi.org/10.3390/bs15050668 . . . . . **157**

Rhys D. Mawson and Sean H. K. Kang
The Distributed Practice Effect on Classroom Learning: A Meta-Analytic Review of Applied Research
Reprinted from: *Behav. Sci.* **2025**, *15*, 771, https://doi.org/10.3390/bs15060771 . . . . . **176**

Marissa K. Hartwig and Doug Rohrer
Students' Perceptions of Effective Math Learning Strategies
Reprinted from: *Behav. Sci.* **2025**, *15*, 1047, https://doi.org/10.3390/bs15081047 . . . . . **204**

Amber E. Witherby, Addison L. Babineau and Sarah K. Tauber
Students', Teachers', and Parents' Knowledge About and Perceptions of Learning Strategies
Reprinted from: *Behav. Sci.* **2025**, *15*, 160, https://doi.org/10.3390/bs15020160 . . . . . **222**

Ezgi M. Yüksel, C. Shawn Green and Haley A. Vlach
Durability of Students' Learning Strategies Use and Beliefs Following a Classroom Intervention
Reprinted from: *Behav. Sci.* **2025**, *15*, 706, https://doi.org/10.3390/bs15050706 . . . . . **246**

Morgan D. Shumaker, Michelle L. Rivers and Sarah K. Tauber
Point Values on Scoring Rubrics Influence Self-Regulated Learning for STEM Material
Reprinted from: *Behav. Sci.* **2025**, *15*, 532, https://doi.org/10.3390/bs15040532 . . . . . **274**

Lindsey Engle Richland and Hongyang Zhao
Relational Reasoning and Educational Applications
Reprinted from: *Behav. Sci.* **2026**, *16*, 210, https://doi.org/10.3390/bs16020210 . . . . . **288**

Faria Sana and Barbara Fenesi
Working Memory and Instructional Fit: Reintroducing Aptitude–Treatment Interaction in Education Research
Reprinted from: *Behav. Sci.* **2025**, *15*, 765, https://doi.org/10.3390/bs15060765 . . . . . **322**

Mark A. McDaniel, Christopher M. Wally, Regina F. Frey and Hayley K. Bates
Individual Differences in Student Learning: A Comparison Between the Student Approaches to Learning and Concept-Building Frameworks
Reprinted from: *Behav. Sci.* **2025**, *15*, 1055, https://doi.org/10.3390/bs15081055 . . . . . **346**

Jeri L. Little and Jexy A. Nepangue
Whether Interleaving or Blocking Is More Effective for Long-Term Learning Depends on One's Learning Strategy
Reprinted from: *Behav. Sci.* **2025**, *15*, 662, https://doi.org/10.3390/bs15050662 . . . . . **370**

Shaohang Liu, Christopher Kent and Josie Briscoe
Monitoring-Based Rewards Enhance Both Learning Performance and Metacognitive Monitoring Accuracy
Reprinted from: *Behav. Sci.* **2025**, *15*, 307, https://doi.org/10.3390/bs15030307 . . . . . **393**

Yan Li
Listen or Read? The Impact of Proficiency and Visual Complexity on Learners' Reliance on Captions
Reprinted from: *Behav. Sci.* **2025**, *15*, 542, https://doi.org/10.3390/bs15040542 . . . . . **408**

Yunfeng Wei, Nicholas C. Soderstrom, Michelle L. Meade and Brandon G. Scott
Metacognition About Collaborative Learning: Students' Beliefs Are Inconsistent with Their Learning Preferences
Reprinted from: *Behav. Sci.* **2024**, *14*, 1104, https://doi.org/10.3390/bs14111104 . . . . . **426**

Eva J. C. Smits, Niklas Wenzel and Anique de Bruin
Investigating the Effectiveness of Self-Regulated, Pomodoro, and Flowtime Break-Taking Techniques Among Students
Reprinted from: *Behav. Sci.* **2025**, *15*, 861, https://doi.org/10.3390/bs15070861 . . . . . **443**

Maribeth M. Trego, Julia S. Soares and Annie S. Ditta
Snap & Write: Examining the Effect of Taking Photos and Notes on Memory for Lecture Content
Reprinted from: *Behav. Sci.* **2025**, *15*, 561, https://doi.org/10.3390/bs15050561 . . . . . **474**

Karen Arcos, Rebecca Covarrubias and Benjamin C. Storm
Examining the Potential Benefits of Affirming Values on Memory for Educational Information
Reprinted from: *Behav. Sci.* **2025**, *15*, 1033, https://doi.org/10.3390/bs15081033 . . . . . **495**

About the Editors

Jeri L. Little

Jeri L. Little is an Associate Professor at the Department of Psychology at California State University, East Bay. Her work uses research methods from cognitive psychology to examine issues relevant to educational practice, with a particular focus on memory, knowledge representation, and metacognition. Her research is focused on understanding how learning is shaped by the ways learners engage with that information and how evidence-based principles from cognitive psychology can be applied to improve learning. Her research has been recognized by the New York Academy of Sciences. Additionally, she has served on the Editorial Board of the *Journal of Applied Research in Memory and Cognition* and is a Fellow of the Psychonomic Society.

Elizabeth Ligon Bjork

Elizabeth Ligon Bjork is a Professor Emerita at the Department of Psychology at the University of California, Los Angeles. A primary focus of her research has been how the principles of learning discovered in the laboratory—such as the benefits of *desirable difficulties*—can be applied to enhance instructional practices in the classroom as well as to improve self-directed learning. She has served as an Editor or Reviewer for several scientific journals, for NIMH, and for NSF; received awards for mentoring and teaching from UCLA and APS, and is an elected member of the Society of Experimental Psychologist and the American Academy of Arts and Sciences.

behavioral sciences

MDPI

Editorial

A Special Issue Introduction: Educational Applications of Cognitive Psychology

Jeri L. Little [1,*] and Elizabeth Ligon Bjork [2]

[1] Department of Psychology, California State University, East Bay, Hayward, CA 94542, USA
[2] Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095, USA; elbjork@psych.ucla.edu
* Correspondence: jeri.little@csueastbay.edu

1. Introduction

How to optimize knowledge retention and transfer is a fundamental question for students and educators. Recently, the "science of learning" has received widespread interest in popular books like *Make It Stick* (P. C. Brown et al., 2014) and *How We Learn* (Carey, 2015). By examining how people process, store, and retrieve information, cognitive psychologists have identified key mechanisms and processes that underlie effective learning and have developed practical strategies for enhancing learning outcomes (e.g., Dunlosky et al., 2013; Mayer, 2008; Roediger & Pyc, 2012). Additionally, cognitive psychologists are interested in learners' beliefs (i.e., metacognition) about how learning occurs and, particularly, under what circumstances learners' judgments of their own learning misalign with their actual learning outcomes. Finally, cognitive psychologists are increasingly interested in individual differences that can influence the effectiveness of these strategies.

Most of the articles within this Special Issue align with the central themes in the science of learning, including desirable difficulties, metacognition, and individual differences. At the same time, several contributions extended beyond these central themes, reflecting the breadth of contemporary research at the intersection of cognitive psychology and education. Such contributions highlight emerging directions for the science of learning.

In this Introduction, we focus on some of the major research findings on how to optimize retention and transfer and what learners know about these processes, highlighting the publications included in this Special Issue. We then offer practical recommendations and provide avenues for future research.

2. Improving Knowledge Retention with "Desirable Difficulties"

One powerful idea in the science of learning is that some strategies that feel difficult to the learner are actually desirable for long-term retention (R. A. Bjork, 1994; R. A. Bjork & Bjork, 2011, 2020; A. Nelson & Eliasz, 2023). Such "desirable difficulties" include testing oneself rather than rereading (Roediger & Karpicke, 2006; Roediger & Butler, 2011), spacing out study sessions rather than massing them (Cepeda et al., 2006, 2008; Dempster, 1989), and interleaving the presentation or study of items from different categories rather than blocking items by category (Brunmair & Richter, 2019; Firth et al., 2021; Kornell & Bjork, 2008a; Pan, 2015). The desirable difficulty framework builds upon Craik and Lockhart's (1972) levels of processing results: deeper processing of to-be-learned material increases the likelihood that it will be remembered later. To illustrate why testing is beneficial, answering a question about to-be-learned information often involves deeper processing than simply looking at it again. To illustrate why spacing is beneficial, when learners are exposed to

an item multiple times, they will likely think about that item more deeply on successive presentations if time has passed since the previous presentation than if it has not, most likely because they should recognize that their memory of the item has dissipated during that interval. Finally, to illustrate why interleaving is beneficial, if learning items from different categories, learners will tend to process the second item from the category more deeply if it follows an item from a different category than if it follows an item from the same category. This increased learning is likely occurring because seeing items from different categories induces a consideration of what is making these items belong to different categories, while also considering the features necessary for membership in a given category. Considering these conditions of learning together, in the cases of testing, spacing, and interleaving, learners may process items more deeply because those items are less accessible in memory. That lower accessibility of the items when eventually retrieved in a test or seen again in a spaced or interleaved manner—while perhaps making those conditions of learning seem more difficult to the learner—is actually making that learning experience more potent and long-lasting (R. A. Bjork & Bjork, 1992; but see Kornell, 2025).[1]

The testing effect and spacing effect are generally robust across participants and contexts (e.g., Carpenter et al., 2022; Dunlosky et al., 2013) and are often used together to promote learning (Carpenter & DeLosh, 2005; Hser & Wickens, 1989; Karpicke & Bauernschmidt, 2011; Karpicke & Roediger, 2007; Landauer & Bjork, 1978; Soderstrom et al., 2016). The popular use of flashcards, for example, makes use of both testing and spacing (Kornell, 2009; Kornell & Bjork, 2008b). In the cases of both testing and spacing, the learning advantage of the "difficult" condition (testing and spacing vs. restudy and massing, respectively) is often influenced by the retention interval (i.e., the time period between the learning task and a final criterion test). Although it is sometimes the case that on an immediate test, the "easy" condition outperforms the "difficult" condition, the benefit of the "difficult" condition is typically present at a delay (e.g., Roediger & Karpicke, 2006). Critically, strong performance for the items presented in the "easy" condition on immediate tests (or simply knowing that one could regurgitate knowledge immediately after learning) is likely a major reason why learners typically overestimate the efficacy of these "undesirable" conditions for long-term learning (R. A. Bjork & Bjork, 2011). Additionally, learners tend to underestimate the amount of forgetting that occurs over time (Koriat et al., 2004), and items that are just above the threshold for successful retrieval will be easily forgotten (i.e., unable to be retrieved, Kornell & Bjork, 2025).

Although, in general, the testing and spacing effects are robust over many populations and contexts, the optimal difficulty for maximizing learning can also depend on other factors, and knowing these factors can increase the probability of learning success (c.f., Jenkins, 1979; McDaniel & Einstein, 1989). Jenkins (1979), for example, has proposed that in addition to the orienting task (e.g., desirable difficulty), three categories of factors can interact to influence memory: (a) learner characteristics, (b) materials, and (c) features of the criterial test. For learner characteristics, the optimal difficulty could depend on the background of the learner because a "desirable" level of difficulty for one person may be too difficult for another. For example, in the case of learning motor skills and sports, blocking is sometimes better than interleaving for novices, while interleaving is often better than blocking for the more advanced learner (Wulf & Shea, 2002). Other features of the task can influence whether a difficulty is desirable. In this Special Issue, Healy et al. (2025) used a variety of factors that introduce difficulties in language learning to examine which difficulties best help learning. In one experiment, they had participants study English-French translations in small or large blocks, with the idea that larger blocks may offer a desirable difficulty because of the increased spacing. Mostly consistent with those predictions, they found that although smaller blocks promoted better performance

https://doi.org/10.3390/bs16050713

faster than did larger blocks during acquisition, that advantage did not persist on the delayed test. Similarly, although practicing pairs in the French–to-English direction sped correct performance as compared to practicing in the English-to-French direction, that advantage of French-to-English practice did not persist on a delayed test. Although Healy et al. did not find an advantage of the "difficult" condition at the delayed test, they did find an interaction between difficulty and retention interval such that the advantage of the "easy" condition was eliminated over time.

2.1. Tests Serve as Tools for Learning, Not Just Assessment

As previously mentioned, the testing effect (or retrieval practice effect, as it is often called) is the finding that performance on a later test is improved by using retrieval or other testing procedures during practice as compared to re-exposure or restudy of the to-be-learned content (Roediger & Karpicke, 2006). This empirical advantage of retrieval has existed in the psychological literature since the early 20th century (Gates, 1917; Spitzer, 1939). Since then, it has been shown with a wide variety of materials, contexts, and age groups (Adesope et al., 2017; Agarwal et al., 2021; Pan & Rickard, 2018; Roediger & Butler, 2011; Rowland, 2014; Yang et al., 2021) and thus has clear implications for education. An enlightening review of the use of retrieval practice in health professions by Serra et al. (2025) appears in this Special Issue.

Although research has clearly shown that, in general, testing tends to be advantageous for learning, research has also shown that a variety of factors can influence the size of that benefit. The format of the test question (e.g., short-answer, multiple-choice) is one such factor. Early research examining format indicated that cued-recall tests (e.g., short-answer) tended to result in larger learning advantages than did recognition tests and attributed that benefit to the retrieval processes afforded by cued-recall testing (Foos & Fisher, 1988). Because it was believed that performance on multiple-choice tests relies on recognition, whereas performance on cued-recall tests requires retrieval, a prevailing idea was that cued-recall tests would usually be better than multiple-choice tests for producing learning. We (E. L. Bjork et al., 2014; Little, 2018; Little & Bjork, 2015; Little et al., 2012, 2019) have argued that although multiple-choice tests can be written to rely on recognition (and less depth of processing), they can also be written to induce recall—namely, when the incorrect alternatives are competitive (plausible) answer choices. For example, Little and Bjork (2015) had participants read passages and then answer multiple-choice questions about the information in those passages that either had competitive incorrect alternatives or non-competitive incorrect alternatives. Then, on a different test administered after a short delay, participants were better able to answer questions about incorrect alternatives when they had previously been competitive alternatives than when they had been non-competitive alternatives. Additionally, Little et al. (2012) showed that multiple-choice tests could be as or more effective than cued-recall tests for retention–both of previously tested and related content–even when feedback was provided on the initial test.

Using Little and Bjork's (2015) materials, Sparck et al. (2016) had participants give confidence-weighted responses for the alternatives (i.e., if participants were unsure of the correct answer, they could indicate which answers they were leaning toward and whether they were leaning closer to one answer or another) and found this procedure to lead to an even larger advantage for the non-tested related information than did answering standard multiple-choice questions. This finding is consistent with the idea that confidence-weighted judgments promote even deeper processing than do simple multiple-choice questions with competitive alternatives. In a similar vein, in this Special Issue, Clark et al. (2025) manipulated whether or not participants needed to justify their answer to multiple-choice questions as a means to examine the role of answer justification in improving the

https://doi.org/10.3390/bs16050713

benefit of the testing effect with multiple-choice questions. They found that justifying answers increased the learning benefits of answering multiple-choice questions (see also Little et al., 2019).

Usually, the testing effect is examined in situations in which retrieval is overt (i.e., participants need to report what they retrieved in the form of answers on an initial test). Sometimes, however, retrieval is not or cannot be overt, and instead, students only retrieve information in their minds (Witherby et al., 2025b). Some past research has shown that whether retrieval is overt or covert has little effect on learning (Higham et al., 2023; Putnam & Roediger, 2013; Smith et al., 2013), but other research has shown overt retrieval to be more effective than covert retrieval (Abel & Roediger, 2018; Jönsson et al., 2014; Kubik et al., 2020; Sumeracki & Castillo, 2022; Tauber et al., 2018). While the above research was mostly conducted in the laboratory, Witherby et al. (2025b) examined the relative benefits of overt versus covert retrieval in 8th-grade science classes. They obtained preliminary evidence that overt retrieval might be more effective than covert retrieval in an educational setting, but cautioned that this result may have been found because learners failed to engage in retrieval processes when—in the covert condition—they did not need to show evidence of having engaged in such efforts.

Testing can also be beneficial when it occurs before learning, as in the form of pretests (Carpenter & Toftness, 2017; Grimaldi & Karpicke, 2012; James & Storm, 2019; Little & Bjork, 2016; Richland et al., 2009). That is, answering questions before reading or listening to a lecture presenting the to-be-learned material can improve the learning of that material when it is later read or presented in a lecture, especially the learning of the pretested information. Previous research has often found larger advantages of post-testing over pretesting (Geller et al., 2017; McDaniel et al., 2011), but sometimes the advantage of pretests match that of posttests (Pan & Sana, 2021). In this Special Issue, James and Storm (2025) investigated the relative benefits of pretests versus posttests using True/False tests and found that both types of tests led to a similar number of correct responses on the final test. For intrusions, however, True/False pretests led to fewer intrusions than did True/False posttests, even when significant feedback was provided on those posttestssuggesting that when True/False questions are used, pretesting may have advantages over posttesting.

Although testing clearly confers direct benefits to learning, testing can also provide indirect effects. Roediger et al. (2011b) have argued that testing can induce learning in a variety of ways, including providing learners with ideas about what they do and do not know and thus helping them to focus future study on information that is less well known, a point to which we will return in the section on Metacognition. Another indirect effect of testing is that testing some content can increase learning on futurto-be-learned content, the so-called forward testing effect (Pastötter & Bäuml, 2014; Yang & Shanks, 2018; Yang et al., 2018; see also early evidence in Thune, 1950). In this Special Issue, Assadipour et al. (2025) examined the forward testing effect using lecture videos. They compared a condition in which participants retrieved information about previously learned content to a control condition, as well as two other conditions that involved retrieval (i.e., semantic generation and autobiographical retrieval). Then, after watching a 15-min lecture video, participants took a test (or did one of the other retrieval activities), and then watched another 15 min lecture that was followed by another test. Participants' performance on the most recently learned content (i.e., information about the second video) was facilitated by the retrieval of the earlier content as compared to the control condition; furthermore, semantic generation and autobiographical retrieval following the first video did not facilitate learning of the second video. Thus, it seems that regular test-taking of information that was learned will help learners to retain that information and will also help learners to retain new information that is subsequently presented.

https://doi.org/10.3390/bs16050713

In summary, testing is a powerful tool for learning because it can be used in a wide variety of ways and helps learning both directly and indirectly. Directly, recalling information makes that information more recallable in the future. Indirectly, retrieving information can have other advantages, including changing the effectiveness of future study, a point to which we return in the Metacognitive section below.

2.2. Study Schedules Influence Learning: Spacing and Interleaving Effects

Another desirable difficulty involves spacing out rather than massing study, the so-called spacing effect (Cepeda et al., 2006, 2008). Like the testing effect, the spacing effect dates back to early empirical records (Ebbinghaus, 1885/1913) and is also robust across a variety of materials, contexts, and types of learners (Carpenter et al., 2012, 2022; Dempster, 1989; Dunlosky et al., 2013; Fishman et al., 1967; Mawson & Kang, 2025; Rea & Modigliani, 1985; Seabrook et al., 2005; Sobel et al., 2011; Toppino & DeMesquita, 1984; Vlach et al., 2008). Several explanations have been proposed for the benefits of spacing. One prominent idea is that spacing introduces forgetting that, in turn, facilitates future learning (R. A. Bjork & Bjork, 1992). Spacing also increases the potential for encoding variability (R. A. Bjork & Allen, 1970), offering the opportunity to create multiple retrieval routes for later recall, a point to which we will return later.

Closely related to the spacing effect is the interleaving effect in which during category learning, for example, members of the different to-be-learned categories (e.g., Categories A, B, and C) are not presented in the more typical blocked manner (i.e., all and only examples of Category A, then all and only examples of Category B, then all and only those of Category C), but rather in an interleaved manner (e.g., one example from Category A with its label; then one from Category C, with its label; then one from Category B, with its label, and so forth until all the examples from each category have been shown). It was largely assumed that blocked presentation of exemplars would be more effective for the learning of each category than an interleaved presentation would be. Kornell and Bjork (2008a) examined this assumption for the learning of the painting styles of different artists and, surprisingly, found that interleaving the presentations of examples of the different artists' paintings led to better learning of each individual artist's style than did blocking the examples of each individual artist's paintings. Interleaving benefits have been shown with a variety of materials (Brunmair & Richter, 2019; Firth et al., 2021; Pan, 2015; Rohrer, 2012; Rohrer & Taylor, 2007). Given this surprising result, many other studies have attempted to determine more specifically why interleaved category learning would be superior to blocked category learning. For example, research by Birnbaum et al. (2013) has shown that the observed interleaving benefit could not be attributed to spacing alone. A prominent idea for why interleaving is beneficial is that it allows learners to focus on differences between items from different categories (i.e., discriminative contrast e.g., Birnbaum et al., 2013; Kang & Pashler, 2012). Other possible (not mutually exclusive) reasons include the idea that interleaving increases attention (Wahlheim et al., 2011) or decreases mind-wandering (Metcalfe & Xu, 2016). Less robust than testing or spacing, the benefit of interleaving seems to depend more on contextual factors than does testing or spacing. For example, the strength of the effect is dependent on factors like the similarity of the materials (Brunmair & Richter, 2019), whether the items follow an articulable rule (Noh et al., 2016), whether participants can take notes and/or use them on the test (Little et al., 2026), and even individual differences in propensity to memorize or to find rules (Little et al., 2025; Little & Nepangue, 2025). One interpretation for why interleaving is sometimes worse for learning than blocking is that sometimes interleaving introduces too much difficulty to be desirable.

In summary, spacing out study is often potent for learning. Interleaving–a variation on spacing–can also be beneficial, but its benefits depend on more factors. Consequently,

research continues to examine these circumstances so that learners and educators can effectively implement these strategies.

2.3. Metacognition: What Learners Know About Their Learning and How It Affects Their Study Choices

In addition to examining how to optimize learning, researchers have also been interested in what learners think about the effectiveness of their learning—that is, their metacognitions about learning. Research on metacognition has practical implications because metacognitive monitoring (i.e., what learners think about their knowledge) often impacts study decisions in the future (e.g., Koriat & Goldsmith, 1996; Little & McDaniel, 2015b; T. O. Nelson, 1996; T. O. Nelson & Leonesio, 1988; T. O. Nelson & Narens, 1990; Son & Schwartz, 2002). If, for example, learners think that they have already learned something well, they may spend less time studying it in the future; or, if they think that they have not yet learned something well, they may spend more time studying it in the future (e.g., T. O. Nelson & Leonesio, 1988). This seemingly obvious relationship, however, can sometimes be complicated by other factors, including, for example, what learners think they can reasonably learn in the available restudy session. When learners have limited time, they may focus their attention most on items of medium difficulty (region of proximal learning; Kornell & Metcalfe, 2006; Metcalfe, 2002; for a review of the relationship between monitoring and control, see Son & Schwartz, 2002).

In one study examining metacognitive monitoring and restudy decisions following practice testing or restudy of passages, Little and McDaniel (2015b) had participants read text passages about regions of the world (i.e., six regions, each divided into sections of Geography, Climate, and People, for a total of 18 sections; eight sentences/facts in each section). Then, the initial reading of the passages was followed by either a retrieval (practice testing) period in which participants retrieved all of the content they could remember or a restudy period. Participants then estimated the number of facts they would be able to recall for each section of text (e.g., Norway: Geography). Then, they were able to re-read a subset of the passages at their own pace, with the timing of each sentence measured. Finally, they took a free-recall test on information from all of the sections. In terms of metacognition, first, participants had more accurate metacognitive estimates in the retrieval condition than in the restudy condition. Second, when they had the chance to re-read the sentences from half of the sections, they spent more time rereading information from the sections for which they had given the lowest estimates of future recall than the sections they had given the highest estimates of future recall. Critically, their rereading strategies–following from their metacognitive estimates–led to larger gains from rereading in the retrieval condition than in the restudy condition. Thus, practice testing helped learning both directly and indirectly. That is, recalling information during the practice test increased retention, but it also changed how learners allocated their study time during the phase in which they re-read the sentences, thus increasing retention for that initially tested content even more. Interestingly, participants had more accurate metacognitive judgments in the retrieval condition than in the restudy condition; most likely because they based their metacognitive judgments on their estimated success of retrieval during the first test, which was a concrete indicator. In the restudy condition—in which they made judgments based on less concrete information (e.g., fluency)—they greatly overestimated how much they would remember and were less well calibrated in terms of what information they knew well and what information they did not.

As indicated by the above pattern of results, while some aspects of metacognition can be accurate, other aspects can be inaccurate. Learners can have higher confidence than warranted following some study activities, as compared to other study activities that are actually more effective. In Kornell and Bjork's (2008a) research examining the efficacy

of interleaved learning, for example, the majority of participants believed that they had learned an artist's style better when paintings by that artist had been blocked during study versus when they had been interleaved with those of other artists, even though the majority of participants had actually learned better in the interleaved condition. In Roediger and Karpicke's (2006) research examining the retrieval practice effect, participants who received relatively more study time during the learning phase predicted better later recall of the text passages than those who had received less study time but more testing time during the learning phase. The results, however, showed that those receiving relatively more testing time during the learning phase outperformed those receiving relatively more study time during the learning phase.

In addition to examining metacognition and performance within a single study, some researchers have examined metacognition more generally through survey research. In these cases, researchers can often obtain more information about learners' beliefs. Several such studies are included in this Special Issue. For example, Hartwig and Rohrer (2025) examined learners' beliefs about learning efficacy in spaced versus massed and interleaved versus blocked learning. They found that although spaced practice was judged as more enjoyable, nearly half of the participants failed to appreciate its effectiveness for learning. Appreciation of interleaving over blocking was worse: participants neither liked it nor thought that it was effective.

Most of the research on metacognition has examined students' perceptions of their learning. Witherby et al. (2025a), however, also examined teachers' and parents' perceptions of learning strategies, finding that students, teachers, and parents held generally accurate knowledge about the benefits of testing and spacing, rating those strategies higher than most others (e.g., highlighting, rereading, massed study). This pattern suggests that knowledge about the science of learning is being disseminated. The question remains, however, as to whether even if students have some correct metacognitive knowledge, do they use it? Yuksel et al. (2025) examined the durability of students' knowledge and their use of learning strategies following an intervention of teaching those strategies. Although learners tended to maintain their knowledge of effective strategies over time, their use of those effective strategies did not increase over time. Thus, although students can know about effective strategies, their implementation of them does not always occur. Implementation can be difficult because of time constraints and the planning needed.

One way to increase learning is to draw learners' attention to important information. Learners may learn how to attend to information based on cues given by instructors such as by explicitly telling students that a given concept is important. Another way to draw attention to important to-be-learned information is to give it higher point values on an assignment (e.g., Knowlton & Castel, 2022; Castel et al., 2007), which might have implications for how to designate point values on rubrics (Shumaker et al., 2025).

Taken together, findings on the testing effect, spacing effect, and interleaving suggest that the primary contribution of the desirable difficulty framework is not simply the identification of specific techniques, but a broader reconceptualization of learning as a process in which short-term performance and long-term retention are often misaligned. Learners' feelings of learning in the short-term affect their beliefs about their learning (i.e., their metacognition), which often leads to errors in such beliefs when it comes to long-term learning. Advancing the field will require further exploration of when and for whom such difficulties are desirable and how to increase the accuracy of learners' metacognition.

3. Promoting Transfer: Using Learning in New Contexts

Although memorization and retention of information is critical, learners need to be able to take what they have learned in one context and apply it to another. Successful

transfer of knowledge is inconsistent (Thorndike, 1906) and remains one of the major challenges in education (Barnett & Ceci, 2002). To transfer knowledge successfully, one needs both the ability to reason about the relationships between examples and to know when to do it (i.e., relational reasoning; Richland & Simms, 2015; Richland & Zhao, 2026).

3.1. Near Versus Far Transfer

One of the reasons that the successful transfer of knowledge is inconsistent is that successful transfer often depends on the similarity between the context in which information was learned and the new context (Thorndike, 1906). The more similar the contexts, the more likely the transfer. For this reason (among others), psychologists distinguish between near and far transfer, with the idea that near transfer evidences retention whereas far transfer evidences conceptual understanding (Barnett & Ceci, 2002). Barnett and Ceci provided a taxonomy for distinguishing between relatively near and relatively far transfer, considering changes in knowledge domain, physical context, temporal context, functional context, social context, and modality. Far transfer is more challenging because it requires the learner to abstract underlying principles from one context and recognize their relevance to another, which lacks overlapping contextual features (Gick & Holyoak, 1980, 1983).

3.2. Promoting Far Transfer

In their well-known paper, Gick and Holyoak (1980) examined analogical transfer. They provided participants with a story about a general invading a fortress with troops of soldiers that had to–owing to mines–attack using multiple small forces coming from different directions (i.e., convergence strategy). Later, participants were asked to solve a problem involving the destruction of a tumor using a laser (i.e., Duncker's radiation problem, Duncker, 1945), which could also be solved using a "convergence" strategy. Most participants failed to see the connection between the general story and the radiation problem because the deep similarity (i.e., convergence solution) was obscured by surface-level dissimilarity (i.e., military vs. medical problem). However, certain factors could be introduced to help participants solve the radiation problem. For example, exposure to multiple stories with the same strategy but different surface-level details promoted the learning of the "convergence" rule and its application (Gick & Holyoak, 1983; see also Catrambone & Holyoak, 1989). Transfer can be successful when learners understand the concept decontextualized from a specific example. Relatedly, children could transfer a rule ("hide using mimicry") to a new context to the extent that they understood the causal reason for the principle rather than having just memorized it (A. L. Brown & Kane, 1988). In this Special Issue, Richland and Zhao (2026) detail the critical role of this type of reasoning for education, arguing that relational reasoning is core to innovation and creativity (see also Markman & Wood, 2009).

Expertise in a domain depends on the ability to transfer knowledge (Chi et al., 1981; Ericsson & Smith, 1991). Experts develop a deep understanding of their domains and can readily see the deep structure without getting distracted by the superficial details. For example, Chi et al. (1981) showed that experts (i.e., graduate students) in physics could readily see problems utilizing the same underlying principle as conceptually similar, even when they lacked superficial similarity, whereas novices (i.e., undergraduate students) failed to do so. Novices were misled by superficial similarities across problems utilizing different underlying principles.

One way to develop a deep understanding of a domain is to think about the content flexibly and integrate it with one's other knowledge. Note-taking can be valuable for conceptual understanding, as long as that note-taking is generative and non-verbatim (c.f., Mueller & Oppenheimer, 2014). Generative note-taking changes how learners process

https://doi.org/10.3390/bs16050713

information, pushing them to decide what is important, to organize their knowledge, and to integrate it with prior knowledge (e.g., Fiorella & Mayer, 2015; Peper & Mayer, 1978). Generative note-taking should be especially beneficial for far transfer. For example, in an experiment on note-taking, Peper and Mayer (1978) found that overall, note-takers did better on far transfer questions than did non-note-takers, although note-taking did not help near transfer. Note-taking advantages also depended on learners' level of background knowledge, showing evidence for the idea that learning depends on an interaction of factors related to the task, materials, participants, and test type.

Although desirable difficulties like testing and spacing are often used to promote simple retention, those conditions can be used to promote transfer as well. For example, Butler (2010) showed that testing was beneficial not just for recall, but also for one's later ability to answer inference questions that tested the concept in a different domain (see also Van Eersel et al., 2016). Aligned with some arguments we have made so far, Butler argued that testing is beneficial for transfer because repeated testing with different questions should promote encoding variability and multiply retrieval routes in memory (c.f., Bower, 1972; Estes, 1955; Goode et al., 2008). That is, multiple retrieval routes can improve the likelihood of retention but can also improve the likelihood of transfer. Chan et al. (2006) showed that answering some questions about a text passage can facilitate recall of other related information, as long as the practice question could conceivably prompt a learner to think about related content when answering that initial question (see also Chan, 2009, 2010). We (Little, 2018; Little & Bjork, 2015; Little et al., 2012, 2019) showed that using multiple-choice questions with competitive alternatives can increase later recall of information pertaining to those alternatives–even though those questions were not specifically asked on the earlier test–because trying to reject the incorrect alternatives would induce a learner to think about why they are wrong and in so doing, recall related content. In fact, Little et al. (2019) used a protocol that involved asking participants to report how they answered multiple-choice trivia questions and found that although participants sometimes relied on recognition or "just knew the answer," they sometimes reported using processes of elimination, with those processes often leading to the explicit retrieval of information pertaining to the incorrect alternatives. The benefit of multiple-choice questions for retention of tested and related information has been shown in the classroom as well (E. L. Bjork et al., 2014).

Spaced learning also introduces the opportunity for the encoding variability and/or study-phase retrieval processes that could multiply retrieval routes (Appleton-Knapp et al., 2005; Johnston & Uhl, 1976; Maddox, 2016). First, encoding variability could induce learners to think about a to-be-learned item in multiple ways. Second, to the extent that a subsequent presentation reminds one of an earlier presentation, learners may retrieve information about that earlier presentation, promoting covert retrieval (c.f., study-phase retrieval; Thios & D'Agostino, 1976). Not mutually exclusive, both theories could help to explain why spacing is beneficial for learning. Similarly, in interleaving, one can consider an example from one category contrasted with an example from another in one instance and in contrast to an example from a third or fourth category in others. Each comparison may induce the learner to think about a particular category in a different way. For example, considering Kornell and Bjork's (2008a) painting materials, following one painting by Lewis with a painting by Cross might lead the learner to think about brushstrokes; following a different painting by Lewis with a painting by Braque may lead the learner to think about color.

4. How Individual Differences Influence Learning

Individuals have their own abilities (e.g., working memory capacity, structure building), background knowledge, and other characteristics (e.g., personality, propensity to

memorize or find rules) that can affect their learning, with those individual differences sometimes interacting with conditions of learning to influence learning outcomes.

4.1. How Working Memory Influences Learning

Working memory is closely linked to attentional control. Working memory capacity is a relatively stable cognitive trait (Oberauer, 2019; Shipstead et al., 2016) and is widely studied as an individual difference measure (Smith-Peirce & Butler, 2025). Although the benefits of testing and spacing tend to be robust across a variety of learners, sometimes individual differences in ability, such as working memory capacity, can affect the efficacy of these strategies (Sana & Fenesi, 2025). In this Special Issue, Sana and Fenesi (2025) reviewed literature assessing how working memory influences the efficacy of desirable difficulties. Although working memory capacity has been shown to modify the efficacy of retrieval practice in some studies (Agarwal et al., 2017; Tse & Pu, 2012), most studies do not find a relationship (Smith-Peirce & Butler, 2025). For spaced repetition, the higher the working memory capacity, the bigger the advantage of a difficult intervening task (Bui et al., 2013), with these results predicting better advantages of larger spacing for high-working-memory-span individuals than for low-working-memory span individuals. Delaney et al. (2017) did not find that working memory capacity influenced the advantage of spacing, however. Further research is needed on this issue. Sana and Fenesi (2025) argue that educators (perhaps using appropriate technology) could adapt sequences of presentations to individual learners.

4.2. How One's Propensity to Memorize Versus Find a Rule Influences Learning

Another individual difference worth examining is the strategy that people use to learn (e.g., memorization vs. rule-finding). As mentioned, experts are more likely to focus on deep similarities and novices on surface-level similarities (Chi et al., 1981), but some research has aimed to uncover whether individuals may differ in their propensities to memorize or figure out rules (Little & McDaniel, 2015a, 2015c; McDaniel et al., 2014; Juslin et al., 2003). That is, although much research in the science of learning has aimed at investigating the conditions of learning that are best for all learners, it is important to realize that learners bring their own propensities to the task. For example, Little and McDaniel (2015a) showed that when given a category learning task amenable to both rule-finding and memorization, some learners try to find a rule, whereas others focus on memorizing, with these different strategies leading to different learning outcomes (i.e., superficial retention vs. far-transfer, respectively). In their task, learners studied letter-strings from different categories (e.g., animal—blasimark, plant—spaksvot). The rule for category membership was that animals ended with "k" and plants ended with "t." Later, they were tasked with classifying new items that looked similar to an old item but different in the last letter (e.g., blasimart). Those self-reporting as memorizers tended to classify the items based on similarity (e.g., blasimart is an animal), whereas those who self-reported as rule-abstractors tended to classify based on the rule (e.g., blasimart is a plant). Differences in lab-based function-learning tasks (i.e., predicting outputs for given inputs) using analogous rule-learning/exemplar-learner distinctions have been shown to predict performance in educational contexts (Frey et al., 2017, 2020; McDaniel et al., 2022).

The distinction between memorization and rule-abstraction seems to align with the idea of superficial learning versus deep learning in the Student Approaches to Learning Framework (SAL; Biggs, 1993; Biggs et al., 2001; Entwistle & McCune, 2004). In this Special Issue, McDaniel et al. (2025) raised the question of whether individual differences in memorization versus rule-abstraction are predictive of self-reported student approach to learning scores (surface or deep, respectively). Participants completed the SAL self-report

https://doi.org/10.3390/bs16050713

survey and a function-learning task. McDaniel et al. (2025) failed to find a relationship between SAL scores and function-learning performance. This may be because the two constructs lack overlap, but it could also be because student metacognition is inaccurate (Dunlosky & Metcalfe, 2008) and/or self-reports represent intended approaches, not what students are actually able to achieve (Frey et al., 2020). Future research is needed to understand the stability of individual differences in memorization versus rule-abstraction and what such differences have to do with related distinctions (e.g., superficial learning propensities vs. deep learning propensities).

We have already mentioned that which difficulties are desirable might depend on features of the learner. But might the desirableness of the difficulties also depend on what learners are trying to learn? Little et al. (2025) had learners complete a category learning task modeled on the one used by Little and McDaniel (2015a). Learners did response-feedback trials for four letter-strings from each of four categories, with category membership depending on the last letter of the item. Additionally, in this research, participants saw items from a given category either blocked or interleaved. Replicating Little and McDaniel (2015a), Little et al. showed that when learners reported trying to memorize, they tended to classify test items based on similarity to trained items, but when learners reported trying to find a rule, they tended to classify later items based on the rule. Furthermore, when learners were trying to memorize (either by their own initiative or because they were instructed to do so), interleaving was better than blocking for their later classification of items based on similarity to trained items, but when learners were trying to find a rule, blocking was better than interleaving for their later classification of items based on the rule. In this Special Issue, Little and Nepangue (2025) replicated this work using an instruction manipulation (i.e., participants were instructed to memorize or to find a rule) and showed the predicted effects (i.e., blocking better than interleaving when instructed to find a rule; interleaving better than blocking when instructed to memorize) and showed that these effects persist across a delay of 48 h. Note, however, that others have shown advantages of spacing even when learning conceptual information that involves rule-learning/transfer (e.g., McDaniel et al., 2013a, in a function-learning task), so a factor that may need to be considered is the difficulty in extrapolating the rule. One possibility is that if the rule is relatively easy to extrapolate, people will likely benefit from spacing, but if the rule is very difficult to extrapolate, people may benefit from blocking.

5. Practical Implications

The articles in this Special Issue address many issues at the intersection of cognitive psychology and educational practice. Although most early research was conducted in the laboratory, there have now been many attempts to examine the principles of effective learning in real-world contexts. In many cases, the benefits of spacing and testing have emerged even in these practical contexts. For example, the advantage of testing (e.g., via quizzes, clicker questions, etc.) tends to be shown in real classrooms (see review by Agarwal et al., 2021; meta-analysis in this Special Issue, Mawson & Kang, 2025; but see Moreira et al., 2019). Agarwal, McDaniel, Roediger, and colleagues have completed many studies on testing in authentic educational contexts (McDaniel et al., 2007a, 2007b, 2013b; McDaniel et al., 2011; Roediger et al., 2011a). Although classroom research is tremendously valuable, it is underrepresented in the literature because it often takes more resources and coordination to conduct than it takes to conduct lab-based research. Additionally, classroom research must often sacrifice control for increased ecological validity and this lessening of control can sometimes reduce or eliminate effects, serving as a disincentive for investigators to perform this type of research. However, the reduction of effects does not necessarily indicate that the principles discussed here do not have ecological validity. To illustrate,

imagine a within-subjects study implemented in a course using authentic materials to examine the efficacy of practice testing as a tool for learning. In an actual course, students will often study on their own to maximize test performance because grades are extrinsically motivating. If the instructor implements practice testing for some information, students may find that they need to study that information less but may study the other non-tested information more before the test, thus raising that baseline (control) performance. Thus, although practice testing helped, the baseline performance also improved. Such extra study is much less likely to happen in a laboratory experiment.

There can also be other challenges to implementing desirable difficulties in real-world situations—either on the part of the instructor or the student (Hertel, 2020). Testing is probably the easiest high-impact practice to implement in educational contexts. Most obviously, testing can include quizzes and exams, but less obviously, testing can include clicker or polling questions (Deslauriers et al., 2011), think-pair-share activities (Kaddoura, 2013), or review games (Iwamoto et al., 2017). Students can also test themselves. The first author of the present paper implements the aforementioned three strategies and often encourages her students to leave class trying to recall at least one thing that they learned.

Spacing and interleaving are more challenging for teachers to implement because they involve teaching information in a non-linear manner or circling back to information. One relatively painless way for instructors to implement spacing is by making time for reviews at the end of each section and making exams and quizzes cumulative (or, if that is too aversive, including extra credit questions from topics earlier in the course). Instructors can also plan to revisit information in order to connect it with new information. The first author likes to do this in her cognitive psychology class, for example, by introducing the concepts multiple times throughout the semester while making different points about a given concept each time it is reviewed and reminding students that they learned something about this concept earlier in the course. Students can implement spacing through planning study sessions throughout the semester and using flashcards for spaced retrieval.

In order to induce transfer, teachers may want to give students many examples of a concept that differ in superficial features, drawing students' attention to the deep structure or concept (Gick & Holyoak, 1983). For example, in their paper, Little and Nepangue (2025) suggest that when discussing context-dependent memory, one might highlight studies on environmental context (e.g., Godden & Baddeley, 1975), emotional context (e.g., Eich & Metcalfe, 1989), and the context of the processes involved during encoding and retrieval (Fisher & Craik, 1977), explicitly stating that although the studies may seem different on the surface (i.e., because they differ a lot in superficial features), they are all evidence for the same underlying process: encoding specificity. (Although most cognitive psychology instructors likely teach encoding specificity with a similar variety of examples, they may find that many of their students get stuck on the superficial details and fail to see the connection between the different studies unless they make the shared concept very explicit.) In their review, Richland and Zhao (2026) suggest, for example, that drawing attention to one-on-one relational correspondences within representations—sometimes with spatially aligned representations—may be crucial for ensuring that the learner recognizes the value of relational reasoning. Likewise, instructors may need to provide cues (e.g., "solve this problem the same way"). Richland and Zhao provide several additional suggestions for promoting relational reasoning, including building on prior knowledge and supporting novices in developing a relational mindset. To build a relational mindset, for example, an instructor could highlight the importance of relational information while explicitly downplaying the importance of object-based information.

As mentioned above, laboratory research is often conducted before classroom research because laboratory research is usually easier to conduct, and the stakes for participants are

https://doi.org/10.3390/bs16050713

much lower in a lab-based context than in a classroom-based context. Some of the research introduced in this Special Issue has implications for learning in real-world contexts, but future research in real educational contexts should be conducted before making clear recommendations about real-world use. For example, Liu et al. (2025) showed that rewarding accurate metacognition increased performance as compared to rewarding performance (and to a control condition). A future study could examine whether rewarding accurate metacognition in the classroom by adding points on learning activities, quizzes, or exams for accurate metacognitive judgments could improve subsequent performance. Relatedly, justifying one's answers in multiple-choice quizzes may improve performance (Clark et al., 2025), so researchers could examine whether answer justification on low-stakes quizzes improves performance on later exams. In short, how these research studies could be applied in classroom contexts is clear, but future research should examine whether the effects persist in those contexts.

6. Future Research Directions

As discussed above, there are many open areas for future research to investigate the application of research in cognitive psychology to educational contexts. A critical next step for the science of learning is to move from establishing the existence of robust effects toward establishing an integrative framework that more clearly specifies how these effects depend upon the nature of the material, task, learners, and perhaps other contextual features (c.f., Jenkins, 1979). Some papers in this Special Issue also expand upon the contextual features we should consider (e.g., demographic variables of participants, how technological advances affect how people learn). The directions highlighted below illustrate how interacting factors shape learning across domains, contexts, and in an educational environment in which technology is evolving very quickly.

Although we have highlighted testing, spacing, and interleaving as difficulties that are desirable for long-term retention and transfer, there are other difficulties that may be desirable for learning, many of which are discussed elsewhere (R. A. Bjork, 1994; R. A. Bjork & Bjork, 2011, 2020). Indeed, even more desirable difficulties are likely worthy of study. In this Special Issue, Li (2025) examined the extent to which language learners rely on captions while watching a video in a second language. Although learning language through movies and television should improve listening skills, relying on captions may lead learners to improve reading skills rather than listening skills (Vandergrift, 2004). Li found that less proficient learners tend to rely on captions more than do more proficient learners. It might be that less reliance on captions would be too difficult for a less proficient learner, but it might be the case that less reliance on captions is a desirable difficulty for all. This is an empirical question for future research.

Most cognitive research we have discussed here focuses on the individual, but some recent work has examined cognition in groups/collaborative cognition (Rajaram & Pereira-Pasarin, 2010; Rajaram et al., 2025). Although collaborative memory has been studied for some time, less well understood is collaborative metacognition. Metacognition during collaborative learning includes more than individual metacognition (Järvenoja et al., 2017; Järvelä et al., 2016). In this Special Issue, Wei et al. (2025) examined students' metacognition about collaborative learning, and they showed that although students prefer group study, they perceive group study to be less effective than individual study. What is still unknown is how these metacognitive judgments relate to actual learning, and, thus remains a question for future research.

In this Special Issue, we have also highlighted the effect of learning strategies on learning outcomes. Other factors, such as motivation, may influence learning outcomes. For example, Smits et al. (2025) examined whether different break techniques influence motiva-

https://doi.org/10.3390/bs16050713

tion, productivity, and fatigue. Although they did not find any group differences between different break techniques on any of the variables they examined, further examination of how breaks influence learning seems an important area for future research.

Additionally, technology is constantly evolving in terms of how information is delivered and stored and is increasingly changing how learners engage during learning. As mentioned by Sana and Fenesi (2025), computerized programs or applications may be able to deliver optimum spacing/interleaving depending on aspects of the materials and the abilities of the learner. Also, the way in which information is presented and saved has changed over time. Laptops enable students to take notes more quickly. Being able to take notes more quickly, however, may have negative consequences for learning because it reduces the need to think deeply about what information one should write down (Mueller & Oppenheimer, 2014). For the last few decades, many professors have given lectures using slides, sometimes providing these slides to their students to prepare for the lecture or to study afterward. Students like having access to the slides, but research on the effectiveness of having the slides for learning is mixed (e.g., Grabe et al., 2005; León & García-Martínez, 2021; Weatherly et al., 2003). Of course, even if professors do not post their slides, students can quickly capture the content in class by taking pictures with their phones. It is important to understand how such practices might influence learning. Previous research has shown that photo-taking can impair memory of the photographed information compared to non-photographed information (Henkel, 2014; Lurie & Westerman, 2021; Niforatos et al., 2017; Soares & Storm, 2018, 2022), but in this Special Issue, we see that such impairment is not always the case (Trego et al., 2025). The rise of artificial intelligence and LLMs (large language models) in the last few years has changed more than just how information is delivered and stored—and it has the potential to drastically change how and what people are learning. A concern is that it will reduce the need to use effortful processing that supports long-term learning. Future research can examine how AI can be leveraged to improve rather than reduce learning.

Finally, more research is needed regarding how individual differences influence learning and how the effects discussed in this Introduction might be different for different groups of people. Recent critiques of reliance on WEIRD samples (i.e., people from Western, educated, industrialized, rich, democratic nations) should be considered (e.g., Rad et al., 2018), and it remains important to examine how these principles apply across more diverse populations. That said, even within the traditional groups studied in research, many potentially important individual differences have mostly been ignored within the cognitive literature (e.g., socio-economic status, race, first- vs. continuing-generation status of a college student). For example, a demographic shift in recent decades involves the presence of more first-generation college students (i.e., students who are the first in their family to attend college). It is important to include demographic differences, such as whether students are first- or continuing-generation students, in order to see how principles of the science of learning play out across these types of different populations (see, e.g., Arcos et al., 2025). Fortunately, there is some evidence that active learning (including some of the methods reviewed in the present paper) is even more beneficial for traditionally underrepresented students (Theobald et al., 2020), although much more research is needed on this issue.

7. Conclusions

The field of cognitive psychology has clear implications for educational practice. The field is vast, and this Special Issue and the present Introduction focused on only a few robust principles. To summarize, study strategies that feel easy for learners are not always desirable in terms of producing learning that is both durable and generalizable, and learners often fail to appreciate the desirability of these difficulties–at least when they are in the

https://doi.org/10.3390/bs16050713

midst of learning (e.g., R. A. Bjork & Bjork, 2011). Research on such desirable difficulties (e.g., testing oneself, spacing out one's study, and interleaving examples from different categories) has applications for how to construct more effective learning experiences. Considering desirable difficulties in light of other factors (i.e., of the individual, the materials, and the test) provides additional factors to consider when applying this knowledge, and considering these factors can increase our ability to predict learning outcomes. This Special Issue goes beyond these principles, however, by including 23 papers addressing a range of research at the intersection of cognitive psychology and education. The included papers provide more specification for the principles above, with replications of previous findings in new contexts, provision of novel findings, synthesis of previous literature (reviews/meta-analyses), and the introduction of new research questions. Advancing the field will require a more nuanced understanding of how principles described here interact with learner characteristics and educational contexts. As research continues to bridge cognitive psychology and education, the principles discussed here have the potential to shape how learning is designed and experienced across contexts, resulting in better teaching and more effective and efficient learning.

Author Contributions: J.L.L. wrote the initial draft of this manuscript. E.L.B. provided valuable feedback. All authors have read and agreed to the published version of the manuscript.

Funding: This work received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

[1] Note that there are many potential reasons for why testing, spacing, and interleaving improve learning. This paper is not meant to provide an exhaustive list of theories, but we will discuss some other theories later in this paper.

References

Abel, M., & Roediger, H. L. (2018). The testing effect in a social setting: Does retrieval practice benefit a listener? *Journal of Experimental Psychology: Applied*, *24*(3), 347–359. [CrossRef] [PubMed]

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. [CrossRef]

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory*, *25*(6), 764–771. [CrossRef]

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*, *33*(4), 1409–1453. [CrossRef]

Appleton-Knapp, S. L., Bjork, R. A., & Wickens, T. D. (2005). Examining the spacing effect in advertising: Encoding variability, retrieval processes, and their interaction. *Journal of Consumer Research*, *32*(2), 266–276. [CrossRef]

Arcos, K., Covarrubias, R., & Storm, B. C. (2025). Examining the potential benefits of affirming values on memory for educational information. *Behavioral Sciences*, *15*(8), 1033. [CrossRef]

Assadipour, Z., Ahn, D., & Chan, J. C. K. (2025). Interpolated retrieval of relevant material, not irrelevant material, enhances new learning of a video lecture in-person and online. *Behavioral Sciences*, *15*(5), 668. [CrossRef]

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. [CrossRef]

Biggs, J. B. (1993). What do inventories of students' learning processes really measure? A theoretical review and clarification. *British Journal of Educational Psychology*, *63*(1), 3–19. [CrossRef]

Biggs, J. B., Kember, D., & Leung, D. Y. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, *71*(1), 133–149. [CrossRef]

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, *41*, 392–402. [CrossRef]

Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition*, *3*(3), 165–170. [CrossRef]

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 567–572. [CrossRef]

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Erlbaum.

Bjork, R. A., & Bjork, E. L. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, *9*(4), 475–479. [CrossRef]

Bower, G. H. (1972). Stimulus sampling theory of encoding variability. In A. W. Melton, & E. Martin (Eds.), *Coding processes in human memory* (pp. 85–123). Wiley.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning from example. *Cognitive Psychology*, *20*, 493–523. [CrossRef]

Brown, P. C., Roediger, H. L., III, & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Harvard University Press.

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. [CrossRef]

Bui, D. C., Maddox, G. B., & Balota, D. A. (2013). The roles of working memory and intervening task difficulty in determining the benefits of repetition. *Psychonomic Bulletin and Review*, *20*(2), 341–347. [CrossRef]

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133. [CrossRef]

Carey, B. (2015). *How we learn: The surprising truth about when, where, and why it happens*. Random House Trade Paperbacks.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*(3), 369–378. [CrossRef]

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *19*(5), 619–636. [CrossRef]

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, *1*(9), 496–511. [CrossRef]

Carpenter, S. K., & Toftness, A. R. (2017). The effect of prequestions on learning from video presentations. *Journal of Applied Research in Memory and Cognition*, *6*(1), 104–109. [CrossRef]

Castel, A. D., Farb, N. A., & Craik, F. I. (2007). Memory for general and specific value information in younger and older adults: Measuring the limits of strategic control. *Memory & Cognition*, *35*(4), 689–700. [CrossRef] [PubMed]

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1147–1156. [CrossRef]

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. [CrossRef]

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–1102. [CrossRef]

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, *61*, 153–170. [CrossRef]

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory*, *18*(1), 49–57. [CrossRef]

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*, 533–571. [CrossRef]

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152. [CrossRef]

Clark, S. A., Rivers, M. L., & Overono, A. L. (2025). The role of answer justification in multiple-choice testing: Effects on performance and metacognitive accuracy. *Behavioral Sciences*, *15*(4), 477. [CrossRef] [PubMed]

Craik, F. I. M., & Lockhard, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. [CrossRef]

Delaney, P. F., Godbole, N. R., Holden, L. R., & Chang, Y. (2017). Working memory capacity and the spacing effect in cued recall. *Memory*, *26*(6), 784–797. [CrossRef]

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review*, *1*(4), 309–330. [CrossRef]

Deslauriers, L., Schelew, E., & Wieman, C. (2011). Improved learning in a large-enrollment physics class. *Science*, *332*(6031), 862–864. [CrossRef]

Duncker, K. (1945). On problem-solving (L. S. Lees, Trans.). *Psychological Monographs*, *58*(5), i-113. [CrossRef]

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition: A textbook for cognitive, educational, life span, and applied psychology*. Sage Publications.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. [CrossRef]

Ebbinghaus, H. (1913). *Memory. A contribution to experimental psychology* (H. A. Ruger, & C. E. Bussenius, Trans.). Teachers College, Columbia University. (Original work published 1885).

Eich, E., & Metcalfe, J. (1989). Mood dependent memory for internal versus external events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(3), 443–455. [CrossRef]

Entwistle, N., & McCune, V. (2004). The conceptual bases of study strategy inventories. *Educational Psychology Review*, *16*(4), 325–345. [CrossRef]

Ericsson, K. A., & Smith, J. (Eds.). (1991). *Toward a general theory of expertise: Prospects and limits*. Cambridge University Press.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review*, *62*(5), 369–377. [CrossRef]

Fiorella, L., & Mayer, R. E. (2015). *Learning as a generative activity*. Cambridge University Press.

Firth, J., Rivers, I., & Boyle, J. (2021). A systematic review of interleaving as a concept learning strategy. *Review of Education*, *9*(2), 642–684. [CrossRef]

Fisher, R. P., & Craik, F. I. M. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(6), 701–711. [CrossRef]

Fishman, E. J., Keller, L., & Atkinson, R. C. (1967). Massed versus distributed practice in computerized spelling. *Journal of Educational Psychology*, *59*, 290–296. Available online: https://escholarship.org/uc/item/1wq7p8h1 (accessed on 10 April 2026). [CrossRef] [PubMed]

Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology*, *80*, 179–183. [CrossRef]

Frey, R. F., Cahill, M. J., & McDaniel, M. A. (2017). Students' concept-building approaches: A novel predictor of success in chemistry courses. *Journal of Chemical Education*, *94*(9), 1185–1194. [CrossRef]

Frey, R. F., McDaniel, M. A., Bunce, D. M., Cahill, M. J., & Perry, M. D. (2020). Using students' concept-building tendencies to better characterize average-performing student learning and problem-solving approaches in general chemistry. *CBE—Life Sciences Education*, *19*(3), ar42. [CrossRef]

Gates, A. I. (1917). Recitation as a factor in memorizing. In R. S. Woodworth (Ed.), *Archives of psychology* (Vol. 6, pp. 1–104). Forgotten Books.

Geller, J., Carpenter, S. K., Lamm, M. H., Rahman, S., Armstrong, P. I., & Coffman, C. R. (2017). Prequestions do not enhance the benefits of retrieval in a STEM classroom. *Cognitive Research: Principles and Implications*, *2*(1), 42. [CrossRef] [PubMed]

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355. [CrossRef]

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1–38. [CrossRef]

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331. [CrossRef]

Goode, M. K., Geraci, L., & Roediger, H. L., III. (2008). Superiority of variable to repeated practice in transfer on anagram solution. *Psychonomic Bulletin & Review*, *15*, 662–666. [CrossRef] [PubMed]

Grabe, M., Christopherson, K., & Douglas, J. (2005). Providing introductory psychology students access to online lecture notes: The relationship of note use to performance and class attendance. *Journal of Educational Technology Systems*, *33*(3), 295–308. [CrossRef]

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505–513. [CrossRef]

Hartwig, M. K., & Rohrer, D. (2025). Students' perceptions of effective math learning strategies. *Behavioral Sciences*, *15*(8), 1047. [CrossRef] [PubMed]

Healy, A. F., Schneider, V. I., & Kole, J. A. (2025). Exploring whether making second-language vocabulary learning difficult enhances retention and transfer. *Behavioral Sciences*, *15*(5), 692. [CrossRef]

Henkel, L. A. (2014). Point-and-shoot memories: The influence of taking photos on memory for a museum tour. *Psychological Science*, *25*(2), 396–402. [CrossRef] [PubMed]

Hertel, P. (2020). Everyday challenges to the practice of desirable difficulties: Introduction to the forum. *Journal of Applied Research in Memory and Cognition*, *9*(4), 425–427. [CrossRef]

Higham, P. A., Fastrich, G. M., Potts, R., Murayama, K., Pickering, J. S., & Hadwin, J. A. (2023). Spaced retrieval practice: Can restudying trump retrieval? *Educational Psychology Review*, *35*, 98. [CrossRef]

Hser, Y., & Wickens, T. D. (1989). The effects of the spacing of test trials and study trials in paired associate learning. *Educational Psychology*, *9*(2), 99–120. [CrossRef]

Iwamoto, D. H., Hargıs, J., Taıtano, E. J., & Vuong, K. (2017). Analyzing the efficacy of the testing effect using KahootTM on student performance. *Turkish Online Journal of Distance Education*, *18*(2), 80–93. [CrossRef]

James, K. K., & Storm, B. C. (2019). Beyond the pretesting effect: What happens to the information that is not pretested? *Journal of Experimental Psychology: Applied*, *25*(4), 576–587. [CrossRef]

James, K. K., & Storm, B. C. (2025). Sorry, am I intruding? Comparing performance and intrusions rates for pretested and posttested information. *Behavioral Sciences*, *15*(8), 1060. [CrossRef]

Järvelä, S., Kirschner, P. A., Järvenoja, H., Malmberg, J., Miller, M., & Laru, J. (2016). Socially shared regulation of learning in CSCL: Understanding and prompting individual- and group-level shared regulatory activities. *International Journal of Computer Supported Collaborative Learning*, *11*(3), 263–280. [CrossRef]

Järvenoja, H., Järvelä, S., & Malmberg, J. (2017). Supporting groups' emotion and motivation regulation during collaborative learning. *Learning and Instruction*, *70*, 101090. [CrossRef]

Jenkins, J. J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L. Cermak, & F. L. M. Craik (Eds.), *Levels of processing in human memory (PLE: Memory)* (pp. 429–446). Psychology Press.

Johnston, W. A., & Uhl, C. N. (1976). The contributions of encoding effort and variability to the spacing effect on free recall. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(2), 153–160. [CrossRef]

Jönsson, F. U., Kubik, V., Larsson Sundqvist, M., Todorov, I., & Jonsson, B. (2014). How crucial is the response format for the testing effect? *Psychological Research*, *78*, 623–633. [CrossRef]

Juslin, P., Olsson, H., & Olsson, A. C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*(1), 133–156. [CrossRef] [PubMed]

Kaddoura, M. (2013). Think pair share: A teaching learning strategy to enhance students' critical thinking. *Educational Research Quarterly*, *36*(4), 3–24.

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*, 97–103. [CrossRef]

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1250–1257. [CrossRef]

Karpicke, J. D., & Roediger, H. L., III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 704–719. [CrossRef]

Knowlton, B. J., & Castel, A. D. (2022). Memory and reward-based learning: A value-directed remembering perspective. *Annual Review of Psychology*, *73*(1), 25–52. [CrossRef]

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, *133*(4), 643–656. [CrossRef] [PubMed]

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*(3), 490–517. [CrossRef] [PubMed]

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*(9), 1297–1317. [CrossRef]

Kornell, N. (2025). Does memory accessibility affect how much we learn from studying? *Behavioral Sciences*, *15*(6), 760. [CrossRef]

Kornell, N., & Bjork, R. A. (2008a). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592. [CrossRef]

Kornell, N., & Bjork, R. A. (2008b). Optimising self-regulated study: The benefits—And costs—Of dropping flashcards. *Memory*, *16*(2), 125–136. [CrossRef]

Kornell, N., & Bjork, R. A. (2025). Why empirical forgetting curves deviate from actual forgetting rates: A distribution model of forgetting. *Behavioral Sciences*, *15*(7), 924. [CrossRef]

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 609–622. [CrossRef]

Kubik, V., Jönsson, F. U., de Jonge, M., & Arshamian, A. (2020). Putting action into testing: Enhanced retrieval benefits long-term retention more than covert retrieval. *Quarterly Journal of Experimental Psychology*, *73*(12), 2093–2105. [CrossRef] [PubMed]

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). Academic Press.

León, S. P., & García-Martínez, I. (2021). Impact of the provision of PowerPoint slides on learning. *Computers & Education*, *173*, 104283. [CrossRef]

Li, Y. (2025). Listen or read? The impact of proficiency and visual complexity on learners' reliance on captions. *Behavioral Sciences*, *15*(4), 542. [CrossRef]

Little, J. L. (2018). The role of multiple-choice tests in increasing access to difficult-to-retrieve information. *Journal of Cognitive Psychology*, *30*(5–6), 520–531. [CrossRef]

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, *43*, 14–26. [CrossRef]

Little, J. L., & Bjork, E. L. (2016). Multiple-choice pretesting potentiates learning of related information. *Memory & Cognition*, *44*(7), 1085–1101. [CrossRef] [PubMed]

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*, 1337–1344. [CrossRef]

Little, J. L., Fealy, J. C., Kobayashi, K., & Roth, S. (2026). How note-taking and note-using affects the benefit of interleaving over blocking. *Memory & Cognition*, *54*, 294–310. [CrossRef]

Little, J. L., Frickey, E. A., & Fung, A. K. (2019). The role of retrieval in answering multiple-choice questions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(8), 1473–1485. [CrossRef]

Little, J. L., & McDaniel, M. A. (2015a). Individual differences in category learning: Memorization versus rule abstraction. *Memory & Cognition*, *43*(2), 283–297. [CrossRef]

Little, J. L., & McDaniel, M. A. (2015b). Metamemory monitoring and control following retrieval practice for text. *Memory & Cognition*, *43*(1), 85–98. [CrossRef]

Little, J. L., & McDaniel, M. A. (2015c). Some learners abstract, others memorize examples: Implications for education. *Translational Issues in Psychological Science*, *1*(2), 158–169. [CrossRef]

Little, J. L., Nepangue, J., & Longares, A. (2025). The optimal sequence for learning can depend on one's strategy: An individual differences approach. *Learning and Individual Differences*, *120*, 102684. [CrossRef]

Little, J. L., & Nepangue, J. A. (2025). Whether interleaving or blocking is more effective for long-term learning depends on one's learning strategy. *Behavioral Sciences*, *15*(5), 662. [CrossRef] [PubMed]

Liu, S., Kent, C., & Briscoe, J. (2025). Monitoring-based rewards enhance both learning performance and metacognitive monitoring accuracy. *Behavioral Sciences*, *15*(3), 307. [CrossRef]

Lurie, R., & Westerman, D. L. (2021). Photo-taking impairs memory on perceptual and conceptual memory tests. *Journal of Applied Research in Memory and Cognition*, *10*(2), 289–297. [CrossRef]

Maddox, G. B. (2016). Understanding the underlying mechanism of the spacing effect in verbal learning: A case for encoding variability and study-phase retrieval. *Journal of Cognitive Psychology*, *28*(6), 684–706. [CrossRef]

Markman, A. B., & Wood, K. L. (Eds.). (2009). *Tools for innovation: The science behind the practical methods that drive new ideas*. Oxford University Press.

Mawson, R. D., & Kang, S. H. K. (2025). The distributed practice effect on classroom learning: A meta-analytic review of applied research. *Behavioral Sciences*, *15*(6), 771. [CrossRef]

Mayer, R. E. (2008). *Learning and instruction* (2nd ed.). Pearson Education.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414. [CrossRef]

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007a). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513. [CrossRef]

McDaniel, M. A., Cahill, M. J., Frey, R. F., Limeri, L. B., & Lemons, P. P. (2022). Learning introductory biology: Students' concept-building approaches predict transfer on biology exams. *CBE—Life Sciences Education*, *21*(4), ar65. [CrossRef]

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*, *143*, 668–693. [CrossRef]

McDaniel, M. A., & Einstein, G. O. (1989). Material-appropriate processing: A contextualist approach to reading and studying strategies. *Educational Psychology Review*, *1*(2), 113–145. [CrossRef]

McDaniel, M. A., Fadler, C. L., & Pashler, H. (2013a). Effects of spaced versus massed training in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5), 1417–1432. [CrossRef] [PubMed]

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007b). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*, 200–206. [CrossRef]

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013b). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*(3), 360–372. [CrossRef]

McDaniel, M. A., Wally, C. M., Frey, R. F., & Bates, H. K. (2025). Individual differences in student learning: A comparison between the student approaches to learning and concept-building frameworks. *Behavioral Sciences*, *15*(8), 1055. [CrossRef] [PubMed]

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, *131*(3), 349–363. [CrossRef]

Metcalfe, J., & Xu, J. (2016). People mind wander more during massed than spaced inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(6), 978–984. [CrossRef]

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: A review of applied research. In *Frontiers in education* (Vol. 4, p. 5). Frontiers Media SA.

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science*, *25*(6), 1159–1168. [CrossRef]

Nelson, A., & Eliasz, K. L. (2023). Desirable difficulty: Theory and application of intentionally challenging learning. *Medical Education*, *57*(2), 123–130. [CrossRef]

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item. *Applied Cognitive Psychology*, *10*, 257–260. [CrossRef]

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect.". *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 676–686. [CrossRef]

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 125–141). Academic Press.

Niforatos, E., Cinel, C., Mack, C. C., Langheinrich, M., & Ward, G. (2017). Can less be more?: Contrasting limited, unlimited, and automatic picture capture for augmenting memory recall. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(2), 1–22. [CrossRef]

Noh, S. M., Yan, V. X., Bjork, R. A., & Maddox, W. T. (2016). Optimal sequencing during category learning: Testing a dual-learning systems perspective. *Cognition*, *155*, 23–29. [CrossRef] [PubMed]

Oberauer, K. (2019). Working memory and attention—A conceptual analysis and review. *Journal of Cognition*, *2*(1), 36. [CrossRef]

Pan, S. C. (2015, August 4). The interleaving effect: Mixing it up boosts learning. *Scientific American*. Available online: https://www.scientificamerican.com/article/the-interleaving-effect-mixing-it-up-boosts-learning/ (accessed on 10 April 2026).

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. [CrossRef]

Pan, S. C., & Sana, F. (2021). Pretesting versus posttesting: Comparing the pedagogical benefits of errorful generation and retrieval practice. *Journal of Experimental Psychology: Applied*, *27*(2), 237–257. [CrossRef] [PubMed]

Pastötter, B., & Bäuml, K. H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology*, *5*, 83305. [CrossRef]

Peper, R. J., & Mayer, R. E. (1978). Note taking as a generative activity. *Journal of Educational Psychology*, *70*(4), 514–522. [CrossRef]

Putnam, A. L., & Roediger, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition*, *41*(1), 36–48. [CrossRef]

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405. [CrossRef]

Rajaram, S., Greeley, G. D., & Peña, T. (2025). Collaborative memory: A selective review. In *Learning and memory: A comprehensive reference* (3rd ed.). Elsevier. [CrossRef]

Rajaram, S., & Pereira-Pasarin, L. P. (2010). Collaborative memory: Cognitive research and theory. *Perspectives on Psychological Science*, *5*(6), 649–663. [CrossRef] [PubMed]

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning: Journal of Practical Research & Applications*, *4*, 11–18. Available online: https://psycnet.apa.org/record/1986-07610-001 (accessed on 10 April 2026).

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. [CrossRef]

Richland, L. E., & Simms, N. (2015). Analogy, higher-order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 177–192. [CrossRef]

Richland, L. E., & Zhao, H. (2026). Relational reasoning and educational applications. *Behavioral Sciences*, *16*(2), 210. [CrossRef]

Roediger, H. L., III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011a). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395. [CrossRef]

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. [CrossRef]

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255. [CrossRef]

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011b). Ten benefits of testing and their applications to educational practice. In *Psychology of learning and motivation* (Vol. 55, pp. 1–36). Elsevier. [CrossRef]

Roediger, H. L., III, & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition*, *1*(4), 242–248. [CrossRef]

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review*, *24*(3), 355–367. [CrossRef]

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*, 481–498. [CrossRef]

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. [CrossRef]

Sana, F., & Fenesi, B. (2025). Working memory and instructional fit: Reintroducing aptitude–treatment interaction in education research. *Behavioral Sciences*, *15*(6), 765. [CrossRef] [PubMed]

Seabrook, R., Brown, G. D. A., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology*, *19*, 107–122. [CrossRef]

Serra, M. J., Kaminske, A. N., Nebel, C., & Coppola, K. M. (2025). The use of retrieval practice in the health professions: A state-of-the-art review. *Behavioral Sciences*, *15*(7), 974. [CrossRef]

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science*, *11*(6), 771–799. [CrossRef]

Shumaker, M. D., Rivers, M. L., & Tauber, S. K. (2025). Point values on scoring rubrics influence self-regulated learning for STEM material. *Behavioral Sciences*, *15*(4), 532. [CrossRef]

Smith, M. A., Roediger, H. L., III, & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1712–1725. [CrossRef]

Smith-Peirce, R. N., & Butler, A. C. (2025). A scoping review of research on individual differences in the testing effect paradigm. *Learning and Individual Differences*, *118*, 102602. [CrossRef]

Smits, E., Wenzel, N., & de Bruin, A. (2025). Investigating the effectiveness of self-regulated, pomodoro, and flowtime break-taking techniques among students. *Behavioral Sciences*, *15*(7), 861. [CrossRef]

Soares, J. S., & Storm, B. C. (2018). Forget in a flash: A further investigation of the photo-taking-impairment effect. *Journal of Applied Research in Memory and Cognition*, *7*(1), 154–160. [CrossRef]

Soares, J. S., & Storm, B. C. (2022). Does taking multiple photos lead to a photo-taking-impairment effect? *Psychonomic Bulletin & Review*, *29*(6), 2211–2218. [CrossRef]

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*(5), 763–767. [CrossRef]

Soderstrom, N. C., Kerr, T. K., & Bjork, R. A. (2016). The critical importance of retrieval—And spacing—For learning. *Psychological Science*, *27*(2), 223–230. [CrossRef] [PubMed]

Son, L. K., & Schwartz, B. L. (2002). The relation between metacognitive monitoring and control. In T. J. Perfect, & B. S. Schwartz (Eds.), *Applied metacognition* (pp. 15–38). Cambridge University Press.

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2016). On the learning benefits of confidence-weighted testing. *Cognitive Research: Principles and Implications*, *1*(1), 3. [CrossRef] [PubMed]

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*(9), 641–656. [CrossRef]

Sumeracki, M. A., & Castillo, J. (2022). Covert and overt retrieval practice in the classroom. *Translational Issues in Psychological Science*, *8*(2), 282–293. [CrossRef]

Tauber, S. K., Witherby, A. E., Dunlosky, J., Rawson, K. A., Putnam, A. L., & Roediger, H. L. (2018). Does covert retrieval benefit learning of key-term definitions? *Journal of Applied Research in Memory and Cognition*, *7*(1), 106–115. [CrossRef]

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., Chambwe, N., Cintrón, D. L., Cooper, J. D., Dunster, G., Grummer, J. A., Hennessey, K., Hsiao, J., Iranon, N., Jones, L., II, Jordt, H., Keller, M., Lacey, M. E., Littlefield, C. E., . . . Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences*, *117*(12), 6476–6483. [CrossRef]

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior*, *15*(5), 529–536. [CrossRef]

Thorndike, E. L. (1906). *Principles of teaching*. Seiler.

Thune, L. E. (1950). The effect of different types of preliminary activities on subsequent learning of paired-associate material. *Journal of Experimental Psychology*, *40*(4), 423–438. [CrossRef] [PubMed]

Toppino, T. C., & DeMesquita, M. (1984). Effects of spacing repetitions on children's memory. *Journal of Experimental Child Psychology*, *37*, 637–648. [CrossRef]

Trego, M. M., Soares, J. S., & Ditta, A. S. (2025). Snap & write: Examining the effect of taking photos and notes on memory for lecture content. *Behavioral Sciences*, *15*(5), 561. [CrossRef]

Tse, C. S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends of trait test anxiety and working-memory capacity. *Journal of Experimental Psychology: Applied*, *18*(3), 253–264. [CrossRef]

Vandergrift, L. (2004). 1. Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, *24*, 3–25. [CrossRef]

Van Eersel, G. G., Verkoeijen, P. P., Povilenaite, M., & Rikers, R. (2016). The testing effect and far transfer: The role of exposure to key information. *Frontiers in Psychology*, *7*, 1977. [CrossRef]

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, *109*(1), 163–167. [CrossRef]

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, *39*, 750–763. [CrossRef]

Weatherly, J. N., Grabe, M., & Arthur, E. I. (2003). Providing introductory psychology students access to lecture slides via Blackboard 5: A negative impact on performance. *Journal of Educational Technology Systems*, *31*(4), 463–474. [CrossRef]

Wei, Y., Soderstrom, N. C., Meade, M., & Scott, B. G. (2025). Metacognition about collaborative learning: Students' beliefs are inconsistent with their learning preferences. *Behavioral Sciences*, *14*(11), 1104. [CrossRef]

Witherby, A. E., Babineau, A. L., & Tauber, S. K. (2025a). Students', teachers', and parents' knowledge about and perceptions of learning strategies. *Behavioral Sciences*, *15*(2), 160. [CrossRef]

Witherby, A. E., Northern, P. E., & Tauber, S. K. (2025b). Does covert retrieval benefit adolescents' learning in 8th grade science classes? *Behavioral Sciences*, *15*(7), 843. [CrossRef]

Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review*, *9*, 185–211. [CrossRef] [PubMed]

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. [CrossRef] [PubMed]

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *NPJ Science of Learning*, *3*(1), 8. [CrossRef] [PubMed]

Yang, C., & Shanks, D. R. (2018). The forward testing effect: Interim testing enhances inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(3), 485–492. [CrossRef]

Yuksel, E. M., Green, C. S., & Vlach, H. A. (2025). Durability of students' learning strategies use and beliefs following a classroom intervention. *Behavioral Sciences*, *15*(5), 706. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Does Memory Accessibility Affect How Much We Learn from Studying?

Nate Kornell

Department of Psychology, Williams College, 18 Hoxsey Street, Williamstown, MA 01267, USA; nkornell@gmail.com or nk2@williams.edu

Abstract

Two experiments were used to test the hypothesis that studying has a greater impact on learning when the information being studied is currently less accessible in memory. This hypothesis aligns with well-established findings like the spacing effect, but it is inconsistent with other evidence. Prior research has not directly tested the causal relationship between accessibility and learning. Two experiments were used to manipulate memory accessibility using semantic priming. The results indicated that differences in accessibility had no impact on learning. Retrieval difficulty and retrieval success also failed to influence learning. It is speculated that changes in accessibility that are specifically due to forgetting, rather than changes in accessibility per se, may have a causal impact on learning efficiency.

Keywords: memory accessibility; learning; retrieval success; spacing

1. Introduction

Memories vary in how accessible they are. As a memory's level of accessibility increases, the memory will tend to be retrieved with greater success and at faster speeds. Many factors affect memory accessibility; for the purposes of this article, the key ones are that semantic priming and studying both increase memory accessibility, while forgetting results in decreased memory accessibility.

Memory accessibility is correlated with learning: studying produces more learning when a memory's accessibility is lower (e.g., R. A. Bjork & Allen, 1970; Cepeda et al., 2006; Cuddy & Jacoby, 1982). The main focus of this investigation was whether accessibility has a causal impact on learning. Based on previous research, it is possible that low accessibility causes increases in the amount that is learned by studying. It is also possible, however, that accessibility does not play a causal role; for example, in past studies, low accessibility has been coupled with more forgetting, and it is possible that forgetting, rather than accessibility, was playing a causal role. In the present study, the following question was asked: when forgetting does not vary between conditions, will differences in accessibility lead to differences in learning?

1.1. Evidence That Accessibility Is Correlated with Learning

Copious evidence has shown a relationship between learning and accessibility. The largest source of evidence is research on the spacing effect: Studies have shown that spacing study trials on the same item apart leads to more learning than massing them together (e.g., Cepeda et al., 2006; Rohrer et al., 2014). As the gap between two learning trials grows larger (i.e., the spacing increases), the accessibility of the memory during the second learning trial

decreases. In other words, spacing decreases accessibility and increases learning efficiency (e.g., Rohrer, 2009; Rohrer & Taylor, 2006).

Other evidence, beyond spacing effects, has also shown a correlation between accessibility and learning. For example, R. A. Bjork and Allen's (1970) participants studied twice and carried out a short activity between the study trials. The researchers found that when the activity caused more memory interference and, thus, more forgetting, the second study trial became more powerful. Cuddy and Jacoby (1982) also found that forgetting helped memory. Storm et al. (2008) found that learning increased when retrieval-induced forgetting caused the accessibility of memories to decrease.

1.2. Why the Hypothesis That Low Accessibility Causes More Learning Needs Further Testing

Although decreased accessibility and increased learning are correlated, they might not be related causally. Correlational evidence is always open to multiple explanations, and in this case, it is possible that a third variable (such as forgetting) played a causal role and accessibility did not. Moreover, past theorizing has explained the relationship between learning and accessibility without relying on accessibility as a causal factor. In the case of the spacing effect, many theories have been proposed to explain the value of spacing (e.g., Glenberg, 1979; Hintzman, 1974; Ross & Landauer, 1978), including stimulus sampling theories and study-phase retrieval theories, as described below. The present research was not designed to test these theories; rather, it was designed to investigate the possibility that variations in accessibility itself—which is a plausible explanation for the spacing effects given the correlation between accessibility and spacing—have a causal impact on learning.

One prominent class of theories is based on stimulus sampling theory (Estes, 1955; Madigan, 1969) and its application to spacing effects (e.g., Melton, 1970). One such theory, the New Theory of Disuse (R. A. Bjork & Bjork, 1992, 2006), claims that as accessibility decreases, there is an increase in the amount of long-term learning that results from studying. However, this theory does not assume that accessibility plays a causal role; it posits that fluctuations in learning conditions (e.g., contextual variability) that happen over time are responsible for the fact that learning increases as spacing increases.

A second class of theories posits that the spacing effect occurs because the second presentation of an item reminds participants of the first and induces a process of study-phase retrieval (Benjamin & Tullis, 2010; Hintzman, 2010; Hintzman et al., 1975; McKinley & Benjamin, 2020; Thios & D'Agostino, 1976; Wahlheim et al., 2014). Spacing makes the retrieval of a previous learning event more difficult (e.g., R. A. Bjork, 1988) and effortful (e.g., Pyc & Rawson, 2009). For example, Pavlik and Anderson (2005) proposed a theory in which lower levels of memory activation at the time of study are associated with more learning. Again, though, this theory does not assume a causal role for accessibility, but rather suggests that forgetting is a necessary ingredient to enhance learning.

To examine the causal relationship between decreased accessibility and increased learning, I conducted two studies that isolated the effects of accessibility by creating conditions that differed in accessibility but not forgetting. The manipulation involved conceptually priming some words (high accessibility) and not others (low accessibility) prior to restudy. Conceptual priming involves presenting a word that is related to the target word in order to increase the accessibility of the target word in memory (e.g., Collins & Loftus, 1975; Hutchison, 2003; Roediger & Challis, 1992; Soler et al., 2015). On primed trials, participants were shown sentences that contained two words that were related to the target word (for example, "Hawaii is a paradise on earth" was used to prime the word "island"). On unprimed trials, the sentence and target word were not related. If higher accessibility causes less learning to occur, then priming an item prior to restudy should

increase accessibility, decrease learning during restudy, and thereby decrease performance on a final test.

1.3. The Importance of Difficulty for Learning

When a memory becomes less accessible, retrieving the memory becomes more difficult. According to the desirable difficulty framework (e.g., R. A. Bjork, 1994; E. L. Bjork & Bjork, 2011), difficulty during learning can make studying more effective and result in more long-term learning (Soderstrom & Bjork, 2015). Examples of desirable difficulties include spaced practice (Cepeda et al., 2008), interleaving (Rohrer et al., 2019), and testing oneself (Roediger & Butler, 2011). But not all forms of difficulty are desirable for learning–for example, multitasking increases difficulty and impairs learning (Sana et al., 2013). This research examines whether retrieval difficulty is a desirable difficulty.

Recent studies have examined the effect of retrieval success on learning. Retrieval success and accessibility are not the same thing, because retrieval success can be affected by factors other than accessibility, such providing hints or allowing more time for retrieval attempts. Accessibility and retrieval success are related, however, because memories with higher accessibility are more likely to be retrieved successfully. Research has shown that unsuccessful retrieval attempts enhance subsequent learning under some conditions (Grimaldi & Karpicke, 2012; Hays et al., 2013; Knight et al., 2012; Kornell et al., 2009; Richland et al., 2009; Yan et al., 2014). More importantly for the present study, successful and unsuccessful retrieval attempts do not seem to produce different amounts of learning (Chan et al., 2024; Kornell et al., 2015; Maddox et al., 2018; Vaughn et al., 2017; for a review see Kornell & Vaughn, 2016). For example, giving participants carefully designed hints while they carry out retrieval practice does lead to more retrieval success during learning, but it does not impact learning as measured on a subsequent final test (Vaughn & Kornell, 2019). If the present studies show that priming does impact retrieval success, but does not impact learning, then they would add to the accumulating evidence that retrieval success does not impact learning.

2. Experiment 1

Prior research has shown that lower accessibility is correlated with more learning. Experiment 1 investigated whether this correlation is based on a causal relationship whereby low accessibility causes an increase in the amount of learning that results from studying.

2.1. Method

Experiment 1 consisted of three phases. First, the participants studied a word list. Then, they restudied the words under two conditions: a given word was either primed and then restudied (high accessibility) or not primed and then restudied (low accessibility). Memory for the words was tested at the end of the experiment. Experiments 1 and 2 were both preregistered at https://osf.io/st4dx/ (accessed on 21 May 2025). The preregistration includes the methods, stopping rule for data collection, predictions, and data analysis plan.

2.1.1. Participants

I conducted a power calculation assuming a within-participant *t*-test with alpha = 0.05, two tails, and a power of 0.90. I assumed an effect size of d = 0.30 because, in a novel paradigm like this, a small estimated effect size seems appropriate (e.g., d = 0.20), but the theory being tested is that accessibility causes the spacing effect, and spacing effects often have large effect sizes, so a slight upward adjustment of the estimated effect size seemed appropriate. The recommended number of participants was 119.

I preregistered a stopping rule for data collection. The stopping rule made data collection more efficient while also allowing for a larger sample to be run when appropriate,

which helps account for uncertainties in the power analysis (see Dienes, 2016). I planned to recruit 100 participants. After collecting the data and excluding participants based on the criteria described below, I analyzed the data using a Bayes Factor analysis to compare the primed condition to the non-primed condition. I used the following stopping rule: if BF_{10} was less than 0.33 or greater than 3.00, I would stop data collection, but if it fell between these values, I would recruit another 50 participants and repeat this process until I had recruited a maximum of 200 participants.

I excluded participants, based on the preregistered analysis plan, for the following reasons: they did not finish the procedure (n = 3), they said they had participated in the study before (n = 6), the correlation between their pleasantness ratings in phase 1 and 2 was below 0.10 (n = 1; this analysis was meant to detect participants who did not pay attention while making the ratings), they said to exclude them because there had been technical problems (n = 2), their accuracy in judging whether sentences made sense was less than 0.70 (n = 6), or they went through the trials too quickly (n = 8). Going through the trials too quickly was defined as having a median response time under two seconds on pleasantness trials (collapsing across phase 1 and phase 2), having a median response time under five seconds on the sentence judgment task, and having a median response time under five seconds on the final test trials. This rule is a deviation from the preregistered exclusion rule, which said that participants would be excluded if any of these three criteria were met, rather than only being excluded if all three were met. However, the preregistered rule was a typing mistake, and the original intention was to exclude participants if all of these conditions were met, not if any were met. Thus, I deviated from the original rule (which would have meant excluding more than half of the participants) and used the intended rule.

Participants were recruited using Prolific and paid USD 3.00. After collecting the first batch of 100 participants' data, the stopping rule was satisfied, so data collection ended.

The final sample included 77 participants. They were an average of 25 years old; 34 were female, 41 were male, 1 was non-binary/other gender, and 1 did not report their gender.

2.1.2. Materials

The stimuli consisted of target words as well as sentences designed to prime them. To create the stimuli, I used the word-pair norms published by Nelson et al. (1998). I selected word pairs such that the forward association strength (i.e., the rate at which people said the target when given the cue) was between 0.2 and 0.5. Among these pairs, I selected pairs for which a given target had at least two cues in the stimulus set. If a target had more than two cues, I removed cues so that there were only two. I then filtered the stimuli so that all of the words (including targets and primes) were nouns. This filtering produced a set of target words that each had two words that could be used to prime them. For example, the word OCEAN was selected along with primes SEA and WAVE. These primes should elicit OCEAN because the forward association strength was 0.46 for SEA–OCEAN and 0.30 for WAVE–OCEAN.

From this stimulus set, I removed cases where a word (target or prime) was used more than once. If two targets were too similar, I removed one of them. I also removed targets if their two primes related to different two meanings of the target (clay–mold, mildew–mold).

Next, I created sentences that would serve as primes during the experiment. Each sentence contained both prime words for a given target. For example, for the target word OCEAN, which was primed by SEA and WAVE, one of the sentences was "The sea wave was huge". I made two versions of each sentence; one version made sense and the other did not. In this example, the sentence that did not make sense was "The sea wave was tree". The two sentences only varied by their last word, so that participants would have to

read the whole sentence before deciding whether or not the sentence made sense. None of the sentences included any of the target words. In all, the stimuli consisted of 102 target words, each of which was accompanied by two sentences, one that made sense and one that did not.

Because I wanted to test participants by showing them the first two letters of each word, I divided the target words into two sets. One consisted of 64 words whose first two letters were unique within the set. These words were used as targets in the experiment. The remaining 38 words, which had the same first two letters as one of the other targets, were not used as targets, but their sentences were presented during the priming phase of the experiment as non-priming sentences.

2.1.3. Procedure

Following a demographics survey and initial instructions, the experiment was made up of three main phases (see Figure 1).

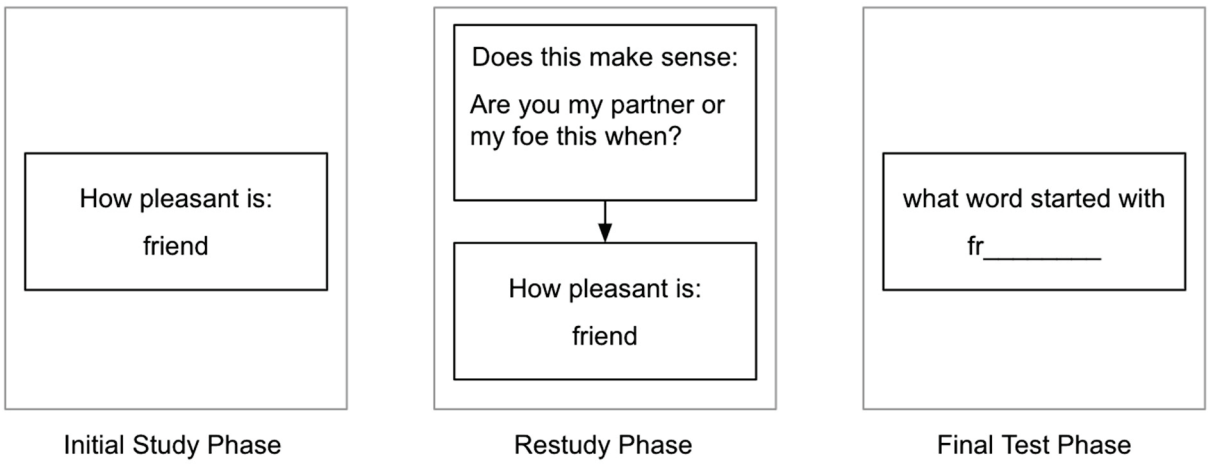


Figure 1. A sample trial from each of the three phases of Experiment 1. Note: During the restudy phase, some sentences made sense and some did not, and some sentences primed (i.e., were related to) the target word and others did not. In this example, the sentence did not make sense but it did prime the target word.

During phase 1, the initial study phase, participants made pleasantness ratings on a 1–5 scale for 48 target words. Each rating was made on a separate trial and participants were given unlimited time to make their ratings. These 48 target words were randomly selected from the 64 possible target words in the stimulus set, and then put in random order, uniquely for each participant.

During phase 2, the restudy phase, there were 48 trials. At the start of each trial, participants were shown a sentence and asked to choose between buttons labeled "makes sense" and "does not make sense". There was no time constraint on responding. After they responded, during the second half of the trial, they were shown a target word and asked to make a pleasantness rating using the same procedure as in phase one.

Phase 2 is when the priming manipulation occurred. Half of the words were primed and half were not. For the 24 words that were primed, the sentence was related to the target word that followed it (e.g., "The sea wave was huge" would be followed by "ocean"). For the 24 words that were not primed, the sentence was not related to the target word (e.g., "Does the dough have butter in it?" would be followed by "school"). In addition, half of the sentences made sense in each condition and half did not. The order in which the words were presented was randomized anew for each participant, with the constraint that words presented in the first half of phase 1 were always presented in the first half of phase 2.

Phase 3, the final test phase, began immediately after phase 2 ended. The participants were tested on all 48 words. On each test trial, the participants were shown the first two letters of a target word and asked to type in the word. There was no time limit on responding. The order of the words was again randomized with the constraint that words presented in the first half of the other two phases were also presented in the first half of this phase.

After the final test phase, the participants were asked whether their data should be excluded because of problems during the study and whether they had taken part in the experiment before.

2.2. Results

On the final test, the accuracy in the Primed and Not Primed conditions was 0.476 (SD = 0.238) and 0.462 (SD = 0.240), respectively (see Figure 2). A Shapiro–Wilk test showed a violation of normality, W = 0.966, $p = 0.036$. A Bayesian Wilcoxon signed rank test showed that $BF_{10} = 0.243$. Because the initial hypothesis was directional, with the prediction that non-primed trials should lead to more learning than primed trials, I conducted a version of the Bayesian Wilcoxon test with the alternative hypothesis of not-primed > primed, and found $BF_{10} = 0.060$. These results support the null hypothesis that accessibility did not impact learning.

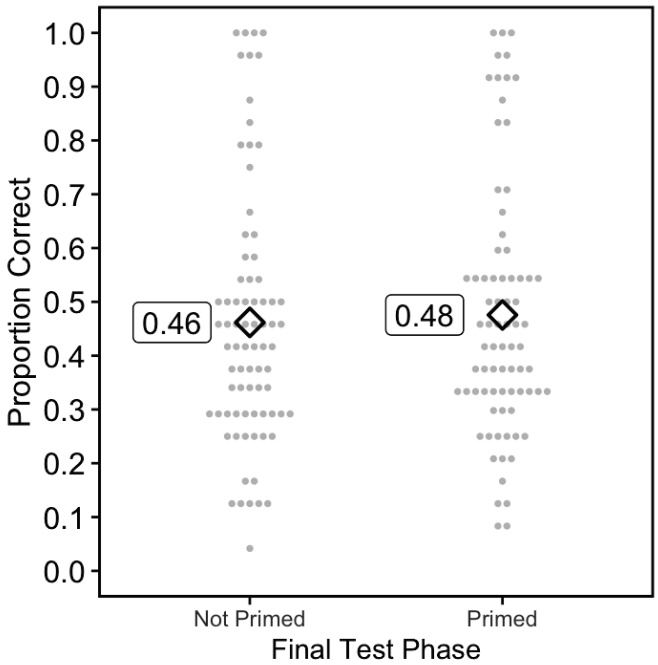


Figure 2. Results of Experiment 1. Note. The diamonds represent the overall means and each circle represents the mean of an individual participant. Each participant was in both conditions (i.e., is represented by two points, one on the left and one on the right).

For completeness, I also carried out a 2 × 2 ANOVA on the final test accuracy, examining two variables: priming and sentence type (i.e., whether sentences made sense or not). The main effects and interaction were not significant. There were no significant main effects of priming ($F(1, 76) = 1.07$, $p = 0.30$, $BF_{incl} = 0.18$) or whether or not the sentences made sense ($F(1, 76) = 0.04$, $p = 0.84$, $BF_{incl} = 0.14$), nor was there a significant interaction ($F(1, 76) = 3.83$, $p = 0.05$, $BF_{incl} = 0.16$). (For more information about the BFincl statistic, see van den Bergh et al., 2020).

When the participants judged whether or not the sentences made sense, they were correct on 0.88 and 0.91 of trials in the Does Not Make Sense and Makes Sense conditions,

respectively. Finally, I examined the pleasantness ratings during the restudy phase. Priming did not affect feelings of pleasantness. The average pleasantness ratings were 3.298 versus 3.298 for Primed and Not Primed (BF_{10} = 0.126).

2.3. Discussion

The results did not support the theory that accessibility is causally linked to learning. Priming a word, which presumably increased its accessibility on the subsequent study trial, did not affect performance on the final test.

Experiment 1 did not measure accessibility during the restudy phase, leaving open the possibility that the sentence primes did not actually influence accessibility during the restudy phase. Experiment 2 was designed to replicate Experiment 1, but also verify that priming impacted accessibility by having participants retrieve the target words from memory during the restudy phase.

3. Experiment 2

The only meaningful difference between Experiment 1 and 2 was that in Experiment 2, during the restudy phase, priming was followed by a retrieval attempt coupled with feedback, rather than a pleasantness trial. The retrieval attempt made it possible to measure memory accessibility during phase 2, and the feedback ensured that participants had a learning opportunity during every trial.

3.1. Method

3.1.1. Participants

Participants were excluded, based on the preregistered analysis plan, for the following reasons: they did not finish the whole study (n = 13), they said they had participated in the study before (n = 9), they said to exclude them because there had been technical problems (n = 6), their accuracy in judging whether sentences made sense was less than 0.75 (n = 29), their median response time on pleasantness trials was less than 1.5 s (n = 9), or their median response time on test trials was less than 3.5 s (n = 2).

The participants were recruited using Prolific and paid USD 3.00. I used the same preregistered stopping rule for data collection as Experiment 1. Following the stopping rule, I collected data from a batch of 100 participants, and then two batches of 50 participants each. The final sample included 145 participants. They were an average of 30 years old; 87 were female, 55 were male, and three were non-binary/other gender.

3.1.2. Procedure

The materials were the same as in Experiment 1. The main procedural change occurred during the restudy phase (see Figure 3). As in Experiment 1, each restudy trial began with a sentence that either did or did not prime the word that followed. In the second half of each trial, instead of restudying via pleasantness trials, the participants were given a two-letter stem (the same stem they would see during the final test) and asked to recall the target word from the initial study phase. After they typed in their answer and pressed enter, they were shown the correct answer for two seconds.

The only other procedural change was that at the end of the study, the participants were asked whether they noticed that the sentences and the word that came after them were sometimes related.

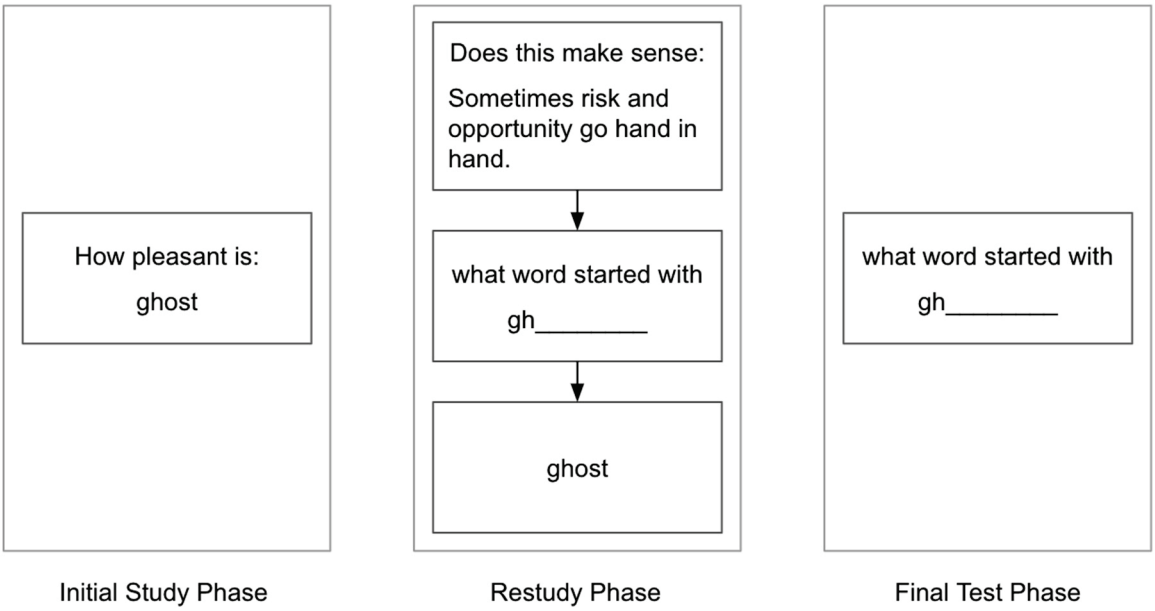


Figure 3. A sample trial from each of the three phases of Experiment 2. Note: During the restudy phase, some sentences made sense and some did not, and some sentences primed (i.e., were related to) the target word and others did not. In this example, the sentence did make sense but it did not prime the target word.

3.2. Results

I first analyzed performance on the test during phase 2 (see Figure 4, left panel). There was a significant effect of priming (BF10 = 5.20 × 1037, d = 1.58). The participants recalled almost twice as many words in the Primed condition (M = 0.613, SD = 0.182) as the Not Primed condition (M = 0.312, SD = 0.174).

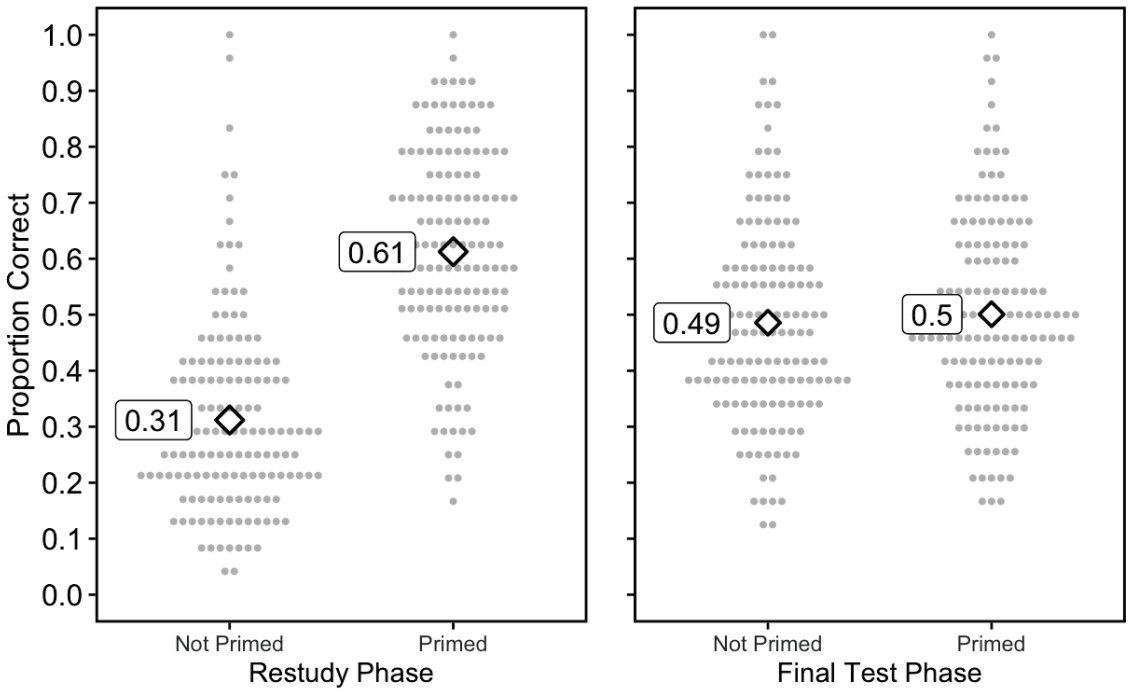


Figure 4. Results of Experiment 2. Note: The diamonds represent the overall means and each circle represents the mean of an individual participant. Each participant was in both conditions in each of the panels (i.e., is represented by two points in each panel, one on the left and one on the right).

During the final test, there was no effect of priming (BF_{10} = 0.22, d = 0.11) (see Figure 4, right panel). The mean proportions correct were 0.501 (SD = 0.176) and 0.486 (SD = 0.185) for the Primed and Not Primed conditions, respectively. Moreover, because the hypothesis was directional, namely, that performance would be higher in the Not Primed condition than the Primed condition, I computed a Bayes Factor using an alternative hypothesis of Not Primed > Primed, and found BF_{10} = 0.041. (Note that unlike Experiment 1, normality was not violated—Shapiro–Wilk W = 0.984, p = 0.080.)

Whether or not the sentences made sense did not affect the accuracy during restudy (BF_{10} = 0.247) or the final test (BF_{10} = 0.097). I also conducted a 2 × 2 ANOVA on final test accuracy, examining two variables: priming and sentence type (i.e., whether sentences made sense or not). The main effects and interaction were not significant. There were no significant main effects of priming (F(1, 144) = 1.76, p = 0.19, BF_{incl} = 0.15) or whether or not the sentences made sense (F(1, 144) = 0.08, p = 0.77, BF_{incl} = 0.08), nor was there a significant interaction (F(1, 144) = 0.001, p = 0.97, BF_{incl} = 0.01).

For completeness, I performed a conditional analysis comparing the final test performance for items that participants had answered correctly during phase 2 (M = 0.68) and items they had not answered correctly (M = 0.33). This difference is likely due to an item-selection effect; participants were more likely to think of an answer on the final test if they had thought of the same answer during the restudy phase.

On the final question, which asked how often the sentence and word were related during the study phase, 78% of participants said about half the time (which is correct), 14% said almost always, and 8% said almost never.

3.3. Discussion

Experiment 2 replicated Experiment 1. Differences in accessibility during study did not impact learning. In addition, Experiment 2 verified that the priming manipulation was successful: accessibility during the restudy phase was higher for words that were primed than for words that were not. Additionally, retrieval success was higher in the primed condition, but the results showed no impact of retrieval success or retrieval difficulty on learning.

It should be noted that accessibility was measured based on a specific memory test–a two-letter stem—and that this measure does not necessarily capture pure semantic activation. Thus, it is an open question whether a different measure of memory would result in differing levels of accessibility during the restudy phase. Similarly, during the test phase, it is possible that another type of memory test would have produced different results. Testing these possibilities is a direction for future research.

4. General Discussion

In two experiments, semantic priming was used to manipulate memory accessibility prior to restudy. The results showed that memory accessibility did not have a causal impact on learning, regardless of whether the restudy involved a presentation (Experiment 1) or retrieval with feedback (Experiment 2).

The correlation between lower memory accessibility and increased learning is well-established (e.g., Cepeda et al., 2006), but the cause of this correlation appears to be a third variable or variables. What might this third variable be? Perhaps certain processes that underlie forgetting cause decreased accessibility and increased learning, creating a correlation between these two variables. This possibility is consistent with the present findings, because the amount of forgetting prior to restudy was the same in the two conditions. It is also consistent with past literature; to my knowledge, in the previous

research showing a correlation between learning and accessibility, the lower accessibility has always been due to increased forgetting.

To be clear, the present experiments cannot definitively support or refute the role of forgetting in learning because they did not examine the effect of forgetting. Nonetheless, multiple theories that explain the correlation between accessibility and learning are consistent with the idea of forgetting as the causal factor, such as theories implicating the benefits of contextual variability, inhibition, or interference (e.g., R. A. Bjork & Bjork, 2006; Cuddy & Jacoby, 1982; Estes, 1955; Storm et al., 2008). Whether forgetting plays a causal role in learning is a question for future research.

These findings suggest that retrieval difficulty, per se, should not be categorized as a desirable difficulty. Desirable difficulties are defined as conditions that decrease performance during study and increase learning (e.g., R. A. Bjork & Bjork, 2006; Soderstrom & Bjork, 2015). In the present experiments, the non-primed conditions decreased performance during study, but this manipulation had no impact on learning. This finding adds to a growing number of studies showing that retrieval success versus failure has little impact on learning (e.g., Kornell & Vaughn, 2016).

Along these lines, although success does not impact learning, making an attempt to think of an answer does (e.g., Knight et al., 2012; Kornell et al., 2009; Yan et al., 2014). It is possible that such attempts contributed to learning in Experiment 2, in both the primed and unprimed conditions. Investigating this possibility is a direction for future research. There are other directions for future research, such as examining the relative contributions of the semantic cue (the sentence) and the two-letter stem. Another question for future research is whether these results would hold under different experimental conditions, such as with other learning materials or longer retention intervals.

Relatedly, theories have been proposed that implicate difficulty as a cause of the spacing effect. According to these theories, studying something a second time can remind learners of the first study trial (e.g., Wahlheim et al., 2014) and induce a process of study-phase retrieval (e.g., Hintzman, 2010). The benefit of spacing occurs because retrieving a previously studied memory has more impact when doing so is more difficult (Benjamin & Tullis, 2010; R. A. Bjork, 1988; Pyc & Rawson, 2009). The present results do not call these theories into question, because these theories explain spacing effects and the present studies did not manipulate spacing. These findings do, however, specify that some ways of manipulating retrieval difficulty may lead to learning gains, but others—including manipulating priming—may not.

5. Concluding Comment

This research attempted to identify causal factors that influence learning. Several variables had no causal impact on learning, including retrieval effort, retrieval difficulty, and memory accessibility. While the literature shows a robust correlation between accessibility and learning, the causes underlying this relationship remain unknown. Future research could investigate causes that are rooted in processes that underlie forgetting to develop a deeper understanding of the causal factors that influence learning.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Williams College (protocol code #2025-16 and date of approval 5 June 2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, *61*(3), 228–247. [CrossRef] [PubMed]

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues, vol. 1: Memory in everyday life* (Vol. 1, pp. 396–401). Wiley. Available online: http://psycnet.apa.org/psycinfo/1988-97682-062 (accessed on 21 May 2025).

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 567–572. [CrossRef]

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healey, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Routledge.

Bjork, R. A., & Bjork, E. L. (2006). Optimizing treatment and instruction: Implications of a new theory of disuse. In L.-G. Nilsson, & N. Ohta (Eds.), *Memory and society: Psychological perspectives* (pp. 116–140). Psychology Press.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. [CrossRef]

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–1102. [CrossRef] [PubMed]

Chan, J. C. K., Davis, S. D., Yurtsever, A., & Myers, S. J. (2024). The magnitude of the testing effect is independent of retrieval practice performance. *Journal of Experimental Psychology: General*, *153*(7), 1816–1837. [CrossRef]

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. [CrossRef]

Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior*, *21*(4), 451–467. [CrossRef]

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. [CrossRef]

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review*, *62*(5), 369–377. [CrossRef]

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition*, *7*(2), 95–112. [CrossRef]

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505–513. [CrossRef]

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *39*(1), 290–296. [CrossRef]

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 77–97). Erlbaum.

Hintzman, D. L. (2010). How does repetition affect memory? Evidence from judgments of recency. *Memory & Cognition*, *38*(1), 102–115. [CrossRef]

Hintzman, D. L., Summers, J. J., & Block, R. A. (1975). Spacing judgments as an index of study-phase retrieval. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(1), 31–40. [CrossRef]

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review*, *10*(4), 785–813. [CrossRef]

Knight, J. B., Ball, H. B., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language*, *66*(4), 731–746. [CrossRef]

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. [CrossRef]

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283–294. [CrossRef]

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. *The Psychology of Learning and Motivation*, *65*, 183–215. [CrossRef]

Maddox, G. B., Pyc, M. A., Kauffman, Z. S., Gatewood, J. D., & Schonhoff, A. M. (2018). Examining the contributions of desirable difficulty and reminding to the spacing effect. *Memory & Cognition*, *46*(8), 1376–1388. [CrossRef]

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior*, *8*(6), 828–835. [CrossRef]

McKinley, G. L., & Benjamin, A. S. (2020). The role of retrieval during study: Evidence of reminding from overt rehearsal. *Journal of Memory and Language*, *114*, 104128. [CrossRef]

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 596–606. [CrossRef]

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*(4), 559–586. [CrossRef]

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. [CrossRef]

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. [CrossRef]

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. [CrossRef]

Roediger, H. L., & Challis, B. H. (1992). Effects of exact repetition and conceptual repetition on free recall and primed word-fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(1), 3–14. [CrossRef]

Rohrer, D. (2009). Avoidance of overlearning characterises the spacing effect. *European Journal of Cognitive Psychology*, *21*(7), 1001–1012. [CrossRef]

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review*, *21*(5), 1323–1330. [CrossRef]

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C.-N. (2019). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology*, *112*(1), 40. [CrossRef]

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology*, *20*(9), 1209–1224. [CrossRef]

Ross, B. H., & Landauer, T. K. (1978). Memory for at least one of two items: Test and failure of several theories of spacing effects. *Journal of Verbal Learning and Verbal Behavior*, *17*(6), 669–680. [CrossRef]

Sana, F., Weston, T., & Cepeda, N. J. (2013). Laptop multitasking hinders classroom learning for both users and nearby peers. *Computers & Education*, *62*, 24–31. [CrossRef]

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*, 176–199. [CrossRef]

Soler, M. J., Dasí, C., & Ruiz, J. C. (2015). Priming in word stem completion tasks: Comparison with previous results in word fragment completion tasks. *Frontiers in Psychology*, *6*, 1172. [CrossRef]

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2008). Accelerated relearning after retrieval-induced forgetting: The benefit of being forgotten. *Journal of Experimental Psychology: Learning Memory and Cognition*, *34*(1), 230–236. [CrossRef] [PubMed]

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior*, *15*(5), 529–536. [CrossRef]

van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E.-J., Derks, K., Dablander, F., Gronau, Q. F., Kucharský, Š., Gupta, A. R. K. N., Sarafoglou, A., Voelkel, J. G., Stefan, A., Ly, A., Hinne, M., Matzke, D., & Wagenmakers, E.-J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année Psychologique*, *120*(1), 73–96. [CrossRef]

Vaughn, K. E., Hausman, H., & Kornell, N. (2017). Retrieval attempts enhance learning regardless of time spent trying to retrieve. *Memory*, *25*(3), 298–316. [CrossRef] [PubMed]

Vaughn, K. E., & Kornell, N. (2019). How to activate students' natural desire to test themselves. *Cognitive Research: Principles and Implications*, *4*(1), 35. [CrossRef] [PubMed]

Wahlheim, C. N., Maddox, G. B., & Jacoby, L. L. (2014). The role of reminding in the effects of spaced repetitions on cued recall: Sufficient but not necessary. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 94–105. [CrossRef]

Yan, V. X., Yu, Y., Garcia, M. A., & Bjork, R. A. (2014). Why does guessing incorrectly enhance, rather than impair, retention? *Memory & Cognition*, *42*(8), 1373–1383. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Why Empirical Forgetting Curves Deviate from Actual Forgetting Rates: A Distribution Model of Forgetting

Nate Kornell [1,*] and Robert A. Bjork [2]

[1] Department of Psychology, Williams College, 18 Hoxsey Street, Williamstown, MA 01267, USA
[2] Department of Psychology, University of California, Los Angeles, CA 90095, USA; rabjork@psych.ucla.edu
* Correspondence: nk2@williams.edu

Abstract

For over a century, forgetting research has shown that recall decreases along a power or exponential function over time. It is tempting to assume that empirical forgetting curves are equivalent to the rate at which individual memories are forgotten. This assumption would be erroneous, because forgetting curves are influenced by an often-neglected factor: the distribution of memory strengths relative to a recall threshold. For example, if memories with normally distributed initial strengths were forgotten at a linear rate, percent correct would not be linear, it would decrease rapidly when the peak of the distribution was crossing the recall threshold and slowly when one of the tails was crossing the threshold. We describe a distribution model of memory that explains the divergence between forgetting curves and item forgetting rates. The model predicts that forgetting curves can be approximately linear (or even concave, like the right side of a frown) when percent correct is high. This prediction was supported by previous evidence and an experiment where participants learned word pairs to a criterion. Beyond its theoretical implications, the distribution model also has implications for education: Creating memories that are just above the threshold helps on short-term tests but does not form lasting memories.

Keywords: forgetting; memory; distribution; threshold; learning

1. Introduction

Previous research has investigated the mathematical function that describes the rate at which memories are forgotten over time (i.e., the forgetting function), dating back to Ebbinghaus (1885). Classic papers have presented empirical data and posited a variety of theories and mathematical functions to describe shapes of forgetting curves (R. B. Anderson, 2001; R. B. Anderson & Tweney, 1997; Averell & Heathcote, 2011; Donkin & Nosofsky, 2012; Loftus, 1985; Wickens, 1998; Wixted, 2004; Wixted & Ebbesen, 1991).

In this article, we take a different approach to understand forgetting curves by emphasizing a factor that is often overlooked in forgetting research: the distribution of memory strengths of the items that are being forgotten (see Figure 1). To do this, we define forgetting in two ways: At a theoretical level, forgetting is a decrease in memory strength, while at the level of empirical observation, it is a change whereby a memory that was previously retrievable can no longer be retrieved. For now, we will define memory strength as how accessible a memory is (we provide a more detailed definition later).

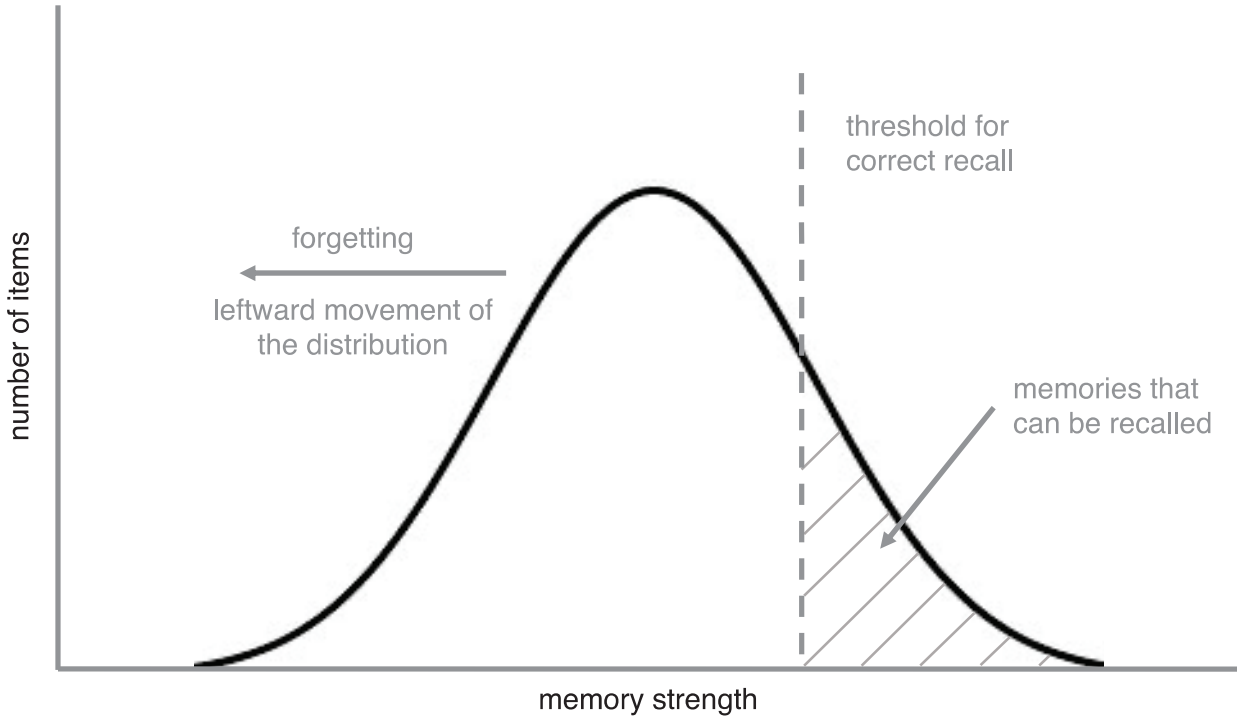


Figure 1. A depiction of forgetting in the distribution model. In this case, a set of memory strengths is normally distributed. Forgetting is represented by the items moving leftward (i.e., losing strength). The percentage of items that are recalled correctly (i.e., that have strengths above the recall threshold) changes as a function of both the rate at which the items move leftward and the shape of the distribution.

Kornell et al. (2011) demonstrated that memory strength distributions can impact empirically observed forgetting rates (this study is described later). Building on this idea, we propose a simple distribution model of forgetting, which posits that these distributions can affect the shape of an empirically observed forgetting curve. Following this simple proposal, we outline predictions of the model, present supporting data, and then sketch outlines of more complicated versions of the model.

The distribution model is not a standalone theory of forgetting, nor is it meant to contradict or replace existing theories. Theories of forgetting such as interference theory (M. Anderson, 2003; Mensink & Raaijmakers, 1988; Underwood, 1957), decay theory (J. Brown, 1958; Hardt et al., 2013), and others (see Oberauer & Lewandowsky, 2008; Wixted, 2004) examine the causes of forgetting at the level of individual items. The distribution model does not focus on individual items but instead takes as given that items are forgotten and adds a second factor, the effect of distributions, to explain empirical forgetting curves. Thus, the distribution model is consistent with existing theories and is an idea that could be added to future theories of forgetting. For these reasons, this article does not provide a comprehensive literature review. The goal of this article is to offer what journalists call an "explainer" (Bahr, 2023): a clear and concise explanation of the distribution model and its relationship to forgetting.

2. An Overview of the Distribution Model

The distribution model makes two key assumptions. The first, which seems self-evident, is that different memories have different strengths. The second is that items are recalled if their memory strength is above a threshold. Thresholds are intrinsic to empirical forgetting curves because the percentage of items recalled is, essentially, the percentage of

items that are above the threshold. Thus, both key assumptions of the distribution model seem applicable to research on forgetting curves.

For an illustration of the distribution model, imagine that you studied 100 word pairs and that each pair had a unique amount of strength in your memory. Further, imagine that these memory strengths fit a normal distribution. Assume that all of these items lost strength (i.e., were forgotten) at the same rate and, to keep things simple (and unrealistic, for now), that the rate of forgetting was linear. Finally, assume that memories are only recalled if they are above a memory strength threshold. Figure 2 shows the percentage of items you would recall in this scenario (i.e., percentage of items above the threshold) as a function of time.

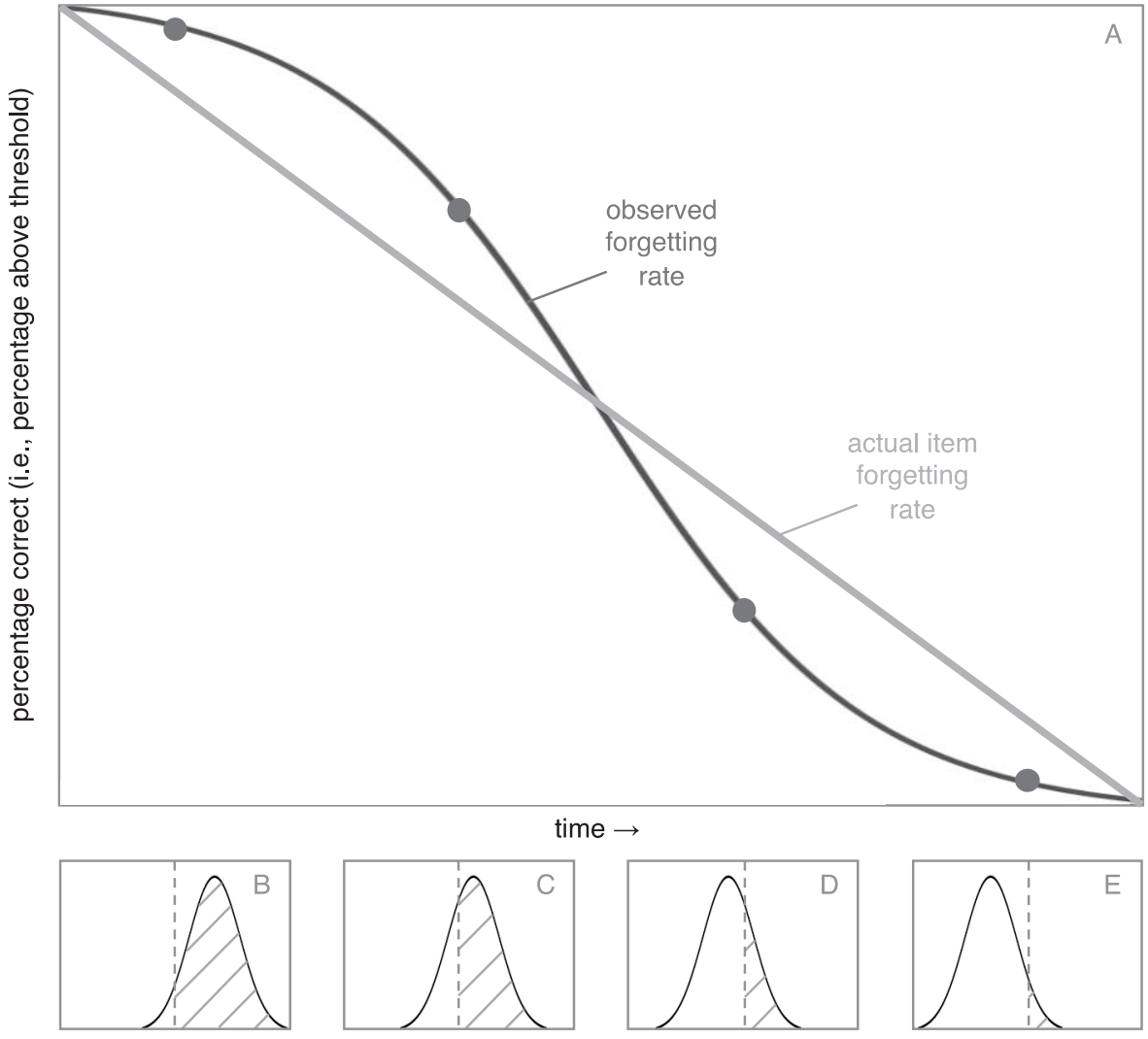


Figure 2. Panel (**A**) depicts the percentage of items above the threshold when a normal distribution moves across a threshold at a constant rate (i.e., when items are forgotten at a linear rate). Panels (**B**–**E**) depict four histograms, with the percentage of items above the threshold (the dashed vertical line) shaded in. Each histogram corresponds to one of the four time points that are identified with dots in panel (**A**). The histogram moves left at a linear rate. In statistical terminology, the proportion of items that are below the threshold in a histogram is referred to as the cumulative distribution function (panel (**A**)). When the underlying distribution is normal, the resulting cumulative distribution function will be s-shaped, which is called a sigmoid function. Thus, panel (**A**) depicts a sigmoid cumulative distribution function, except that it shows the percentage of items above the threshold, rather than below the threshold.

The item forgetting rate is linear, so why is the observed forgetting rate non-linear? The answer lies in the four histograms at the bottom of Figure 2. At first (panel B), relatively few items cross the threshold, because there are few items at the tail of the distribution. The observed forgetting is determined by the number of items that cross the threshold from remembered to forgotten. Thus, there is a relatively low rate of observed forgetting. Then, as the middle of the distribution crosses the threshold, more items are forgotten per unit of time (e.g., between panels C and D). Finally, the observed forgetting returns to a low rate (panel E). The point of Figure 2 is that there are systematic differences between underlying forgetting rates (in this case, linear) and empirically observed forgetting rates (in this case, non-linear).

What happens when we discard the unrealistic idea of a linear forgetting function? Figure 3 makes the more realistic assumption that the forgetting rate of individual items fits a power function (although we are not advocating that power functions are more or less realistic than exponential functions). As Figure 3 shows, the distribution model predicts that the observed forgetting rate (i.e., percent correct) will differ from the item forgetting rate.

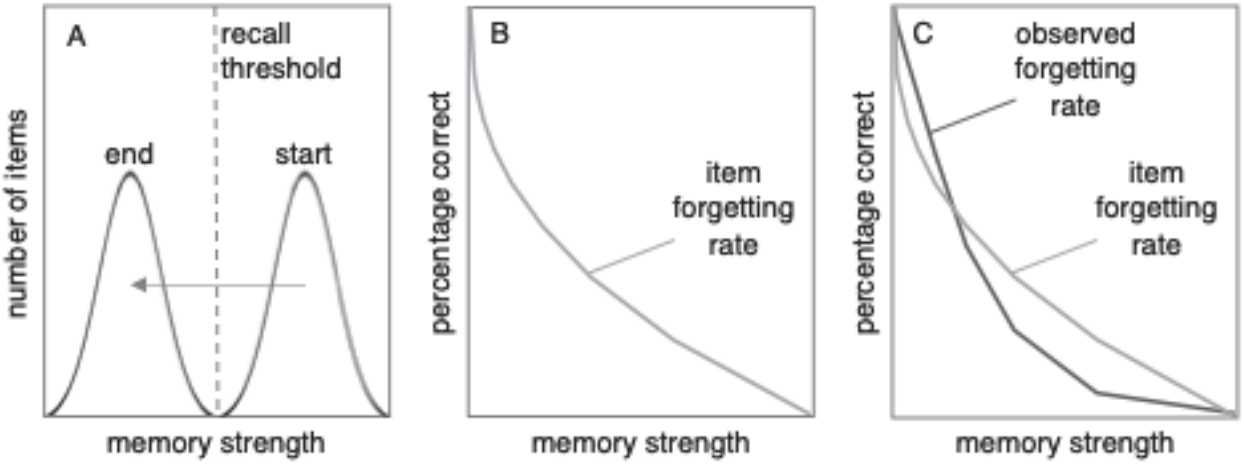


Figure 3. According to the distribution model, an empirically observed forgetting curve (panel (**C**)) is influenced by the distribution of memory strengths (panel (**A**)) and the rate at which individual items are forgotten (panel (**B**)). The item forgetting rate modeled in panels (**B**,**C**) is a power function. The observed forgetting rate in panel (**C**) was created using this function as the item forgetting rate in a simulation of the distribution model.

3. Predictions of the Distribution Model

The distribution model makes testable predictions. Next, we describe three predictions of the model: The observed forgetting rates should be impacted by (a) the section of the distribution being observed, (b) the shape of the distribution, and (c) the standard deviation of the distribution.

3.1. Predictions Based on the Section of the Distribution Being Observed

According to the distribution model, the observed forgetting curve should depend on what slice of the distribution is being captured in the data. If overall performance is high—for example, if percent correct ranges from 50% to 95%—the distribution model predicts that compared to the item forgetting rate, the observed forgetting rate should become flatter and closer to linear (or even progress beyond linear and switch from being convex to concave, as we explain later). The reasoning behind this prediction can be seen in Figures 2 and 3. In the left half of Figure 2, the shape of the distribution drives the curve

to be concave (like the right side of a frown). At the same time, as shown in Figure 3, being forgotten along a power function drives the curve to be convex (like the left half of a smile). These concave and convex forces tend to neutralize each other, leading to a curve that tends toward linearity, as shown in Figure 4, in the area to the left of the dashed vertical line (note: Figure 4 is an enlarged version of Figure 3C). In short, at high rates of recall, the effect of the distribution is to make the observed forgetting curve closer to linear than the underlying item forgetting curve.

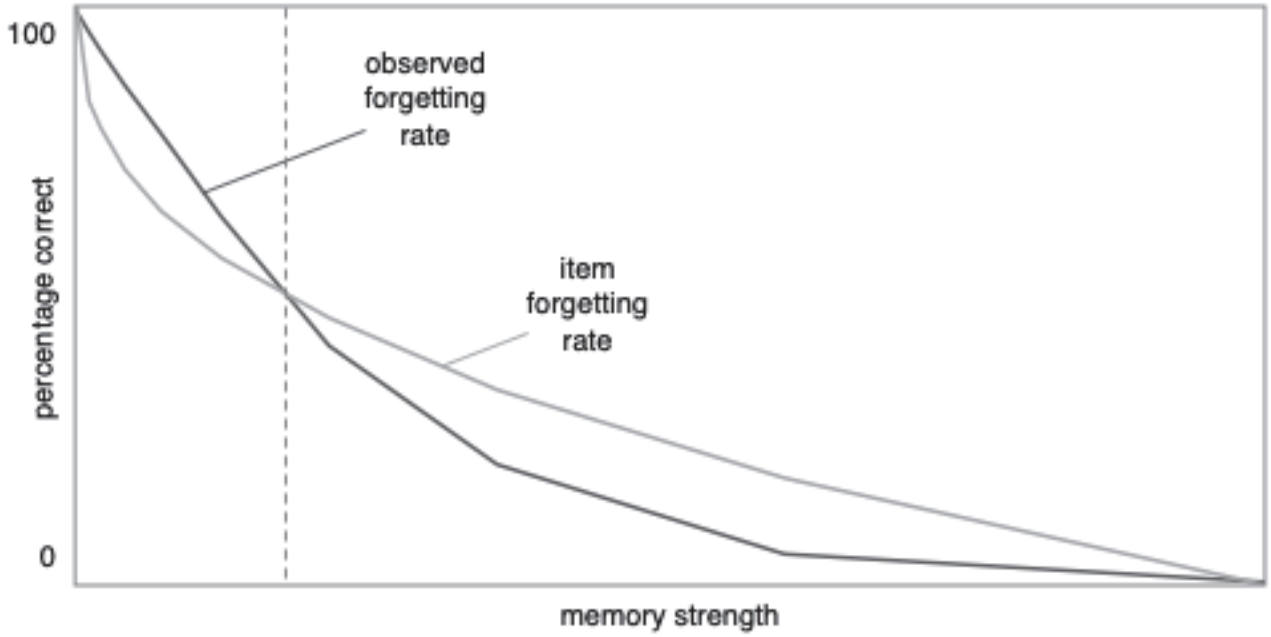


Figure 4. When percent correct is relatively high (to the left of the dashed line), observed forgetting is closer to linear. When percent correct is relatively low, the opposite occurs; observed forgetting is curvier (i.e., has a smaller radius) than the item forgetting curve. This figure was created by enlarging Figure 3C.

If overall performance levels are low (e.g., below 50%), however, the radius of the forgetting curve should decrease (i.e., the curve should become curvier). When the recall performance is low, the distribution drives the curve to be like the left side of a smile (see the right half of Figure 2), and so does the power function. Therefore, the distribution amplifies the effect of the item forgetting rate's power function, as shown in Figure 4 in the area to the right of the dashed line. Thus, at low levels of performance, the observed forgetting curve is more curved (i.e., has a smaller radius) than the item forgetting curve.

Empirical support for these predictions comes from an article called "Linear Forgetting" by Fisher and Radvansky (2019). They conducted three studies and reviewed many additional past studies, all showing linear forgetting. They conclude that "an increased degree of learning appears to be important for linear forgetting to occur" (p. 18). This finding is what the distribution model would predict. (Note that Fisher and Radvansky say, "higher degrees of learning and meaningfully complex materials may be jointly needed to observe linear forgetting." Predictions about complexity, although important, are beyond the scope of this article. Also, note that other studies have not found a relationship between the rate of forgetting and the degree of initial learning; see Rivera-Lares et al., 2022, 2023).

3.2. Predictions Based on the Shape of the Distribution

The distribution model predicts that, all else being equal, the shape of a distribution should affect the empirically observed forgetting curve. This prediction was tested by

Kornell et al. (2011) in a study that examined forgetting rates when underlying distributions had different shapes. Their research compared two conditions. In one condition, participants studied a set of items and presumably developed a roughly normal (or at least unimodal) distribution of memory strengths. In the other condition, participants were tested without feedback. In this case, the items that were recalled correctly, which were already above the recall threshold by virtue of being recalled, moved farther above the threshold because of retrieval practice effects (e.g., Rowland, 2014). The weaker items, which were not retrieved and thus were already below the threshold, were not strengthened because following retrieval failure, the correct answer was not revealed. This second condition, therefore, produced a bifurcated distribution with some items well above the recall threshold and some below it.

The distribution model predicts that a bifurcated distribution like this should appear to be forgotten relatively slowly compared to a unimodal distribution. Forgetting happens when items cross the threshold, and in these bifurcated distributions, the items that started out above the threshold were strong enough that they were well above the threshold. It should take longer for such items to start crossing the threshold, which translates to slower forgetting. By contrast, the other condition, with its normal distribution, should be forgotten more quickly because there were items just slightly above the threshold, and these items should cross the threshold relatively quickly. This prediction of the distribution model was supported by the data.

Furthermore, in Kornell et al.'s (2011) research, there was an additional condition in which participants were tested and given feedback. This condition should not result in a bifurcated distribution because all items could be learned (i.e., potentially move above the threshold) via feedback even if they were not recalled during the study phase (Retrieval attempts followed by feedback enhance learning to roughly the same degree as successful retrieval attempts. See Grimaldi & Karpicke, 2012; Knight et al., 2012; Kornell et al., 2009). In this condition, the distribution model predicts that the observed amount of forgetting should be similar in the testing and study-only conditions, because in both cases the underlying item distribution is expected to be unimodal and roughly normal. This prediction was also supported (Note that the tested items should be farther above the threshold than the studied items, which should affect the observed forgetting rate as shown in Figure 2; however, because there was not a large difference in the mean accuracy scores and thus the positions of the distributions, this effect should be relatively small). In short, these findings showed that the shape of item distributions impacted the rate at which items appeared to be forgotten as predicted by the distribution model.

On this topic, it should be noted that the distribution model is not contingent on the shape of the distribution being normal, or any other shape. The distribution of item strengths should impact the observed forgetting rate as long as different areas of the distribution contain different numbers of items (e.g., there are peaks and low points in the distribution). This claim is related to the topic we discuss next.

3.3. Predictions Based on the Standard Deviation of Distribution

The distribution model predicts that item distributions will have a larger effect on percent correct when the standard deviation of the distribution is smaller. As an extreme example, if the standard deviation was very small and all of the items had roughly the same strength, then they could all pass the threshold at basically the same time. In this case, the empirically observed forgetting curve would look like a step function; it would sit at 100% correct until the item distribution passed the threshold, at which point it would plummet to 0% correct. This would be predicted even if the underlying item forgetting rate was a power function (or any other function).

At the other extreme, when the standard deviation is large, the distribution would behave similarly to uniform, or rectangular, distribution. In other words, there would be about the same number of items at any given point in the distribution. In this case, the distribution's effect would be minimal, and the observed forgetting curve would look roughly the same as the underlying item forgetting curve. In short, a distribution with a smaller standard deviation should have a larger effect on the observed forgetting curve.

There is no strong evidence in support of these predictions. However, indirect support comes from the studies reviewed by Fisher and Radvansky (2019). A common theme among studies that showed linear forgetting was that participants learned to a criterion (i.e., stopped studying after answering correctly a set number of times). Learning to criterion should decrease the standard deviation of item strengths. To illustrate the effect of learning to criterion, assume participants study a set of items to a criterion of three correct responses. When they get an item correct for the third time, it has reached a certain level of strength. At this point, it is no longer studied. As a result, all items reach a similar level of strength. Thus, learning to criterion should decrease the standard deviation of a distribution of memory strengths by ensuring a similar level of learning for all items. One reason this evidence does not provide strong support for the prediction, however, is that learning to criterion simultaneously moves items above the threshold and it decreases their standard deviation. Because these two changes are confounded, it is difficult to know which one impacts empirical forgetting curves. Stronger conclusions could be drawn if future research can unconfound these variables (for example, create a low standard deviation without moving items far above the threshold).

In summary, learning to criterion should make the standard deviation relatively small, which should cause the shape of the distribution to have a relatively larger impact on the observed forgetting curve. Thus, the distribution model predicts that, all else being equal, learning to criterion paradigms should be especially likely to produce linear forgetting for two reasons: the shape of the distribution should have a pronounced effect and, because learning to criterion produces relatively high levels of initial recall, this effect should flatten the distribution (see the portion of Figure 4 to the left of the dashed line). This prediction appears to be consistent with the data reviewed by Fisher and Radvansky (2019). Next, we describe a study that further tested this prediction.

4. Experimental Investigation

We conducted an experiment to test a prediction of the distribution model: The typical convex forgetting curve can go beyond linear and become concave. That is, instead of looking like the left side of a smile, or even a straight line, it might look like the right side of a frown. This prediction does not emerge naturally from models that assume forgetting follows a power or exponential function. Thus, if the data fit a concave curve, it would support the distribution model.

In designing the study, we attempted to create conditions that would take advantage of two predictions we described earlier. (1) In a distribution with a high level of memory strength, the effect of the distribution should be to drive the percent-correct curve toward being a frown shape. (2) When a distribution of items has a small standard deviation, it should amplify the effect of the distribution. We created a high level of memory strength and a small standard deviation by only testing participants on items they had answered correctly during the study phase. The retention interval between studying and testing served as the independent variable.

4.1. Method

Participants studied a set of word pairs, took an initial test, and then took a final test. The retention interval between the initial and final test was 2, 20, or 80 other items. The study's method and data analysis plan were pre-registered at https://osf.io/fpx6d/ (accessed on 21 May 2025).

4.2. Participants

The final sample consisted of 99 participants. Additional participants were removed from the analyses, in accordance with the pre-registration, because they indicated that they had done the experiment before (four participants) or because they left more than half of the answers blank on the final test (two participants). Fourteen additional participants answered zero questions correctly during restudy in at least one condition; because we only analyzed items that were answered correctly during restudy, these participants had no data for at least one condition and were excluded from the analyses.

Participants were paid $2.00. They were recruited using Amazon Mechanical Turk during late January and early February of 2020. Their average age was 38 (SD = 10.9, range = 23–71); 53 reported being female, 45 reported being male, and one did not report their gender. All participants reported that they lived in the United States.

4.3. Materials

The materials were 55 unrelated word pairs such as "sunset-pudding." The words were taken from Paivio et al. (1968). Words were selected if they had an imagery score of 6 or 7 (on a 1–7 scale), a concreteness score of 6 or 7 (on a 1–7 scale), and contained between four and twelve letters. After 110 words were selected, they were paired randomly.

4.4. Procedure

The study was conducted online. Participants encountered each word pair three times during the experiment. The first time was a copy trial. The full pair was shown (e.g., trumpet-house) and below it, the cue word was shown again (e.g., trumpet) and participants were asked to type in the target (e.g., house).

After six intervening trials, participants encountered the pair again. This time, they were given a test trial: They were shown the cue word and asked to type in the target. They were not shown the target after making their response.

The third encounter with a given word pair was another test, again with no feedback. Between the second and third trial, there was a retention interval of 2, 20, or 80 other trials.

A total of 30 pairs were included in the data analysis, with ten pairs in each retention interval. To assign pairs to conditions, a unique randomization was used for each participant.

An additional 25 pairs were shown during the study. Four of these pairs were used during an initial practice phase that consisted of four copy trials followed by four test trials on the same items. Another 21 pairs were used in filler trials that had to be mixed into the other trials in order to achieve the desired retention intervals.

5. Results

In accordance with the pre-registration, items were only analyzed if participants answered them correctly on the initial test. Thirty-three percent of items were excluded. This exclusion alone should not produce concave forgetting, unless the distribution model's predictions are supported. It also had two desired effects: (1) It ensured that all items were above the threshold at the start of the retention interval, and (2) it decreased the standard deviation of item strengths by excluding the weaker items, leaving only a smaller, more tightly clustered set of stronger items.

We predicted that the rate of forgetting would be concave; that is, it would become steeper with time. This prediction was supported; the rate of forgetting (i.e., the slope of the line in Figure 5) was steeper for the second interval, between time 2 and 3, than it was for the first interval, between time 1 and 2. On the final test, the proportion correct was $M = 0.958$, $SD = 0.110$ for time 1; $M = 0.913$, $SD = 0.158$ for time 2; and $M = 0.658$, $SD = 0.298$ for time 3.

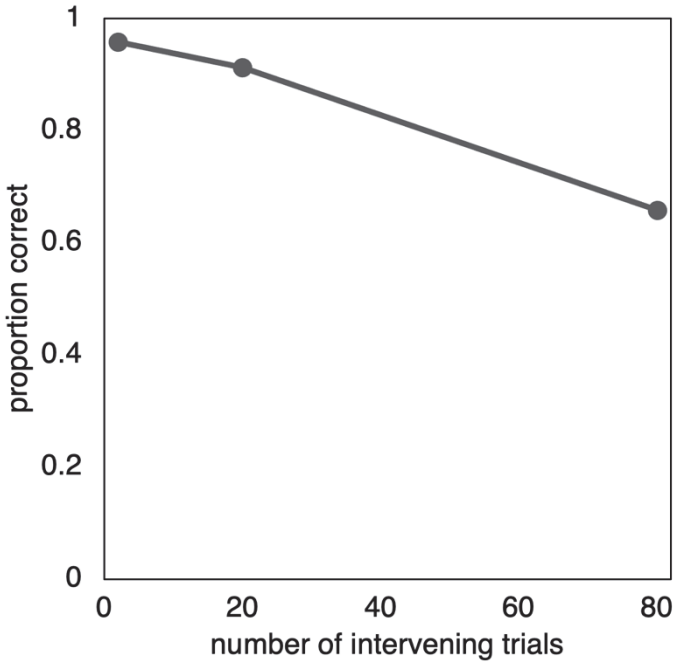


Figure 5. Proportion correct curve showing a concave pattern. Most forgetting curves are convex.

To test whether the increase in slope was significant, we computed forgetting per unit of time for interval 1 and interval 2 for each participant. Forgetting per unit of time is equivalent to the slope of the line. For example, if a participant's percent correct scores were 95, 90, and 70 percent, then the amount forgotten for the two intervals would be $95 - 90 = 5$ items and $90 - 70 = 20$ items, but the two forgetting per unit of time scores would be $(95 - 90)/(20 - 2) = 0.28$ and $(90 - 65)/(80 - 20) = 0.42$. In these equations, the denominator is the number of items separating the two tests; for example, the first and second retention intervals were 2 and 20 items, so we divided by the difference, which is 18 items (To avoid confusion, we should point out that participants were not tested on the same item twice).

The forgetting per unit of time was closer to flat for the first interval ($M = -0.254$, $SD = 0.877$) than the second interval ($M = -0.424$, $SD = 0.481$). When comparing these slopes, a Shapiro–Wilk normality test showed a significant violation of normality ($W = 0.0913$, $p < 0.001$). Therefore, instead of a paired t-test, we compared the scores using a Wilcoxon W test. The difference was significant: $W = 1257$, $p = 0.019$, and rank biserial correlation = 0.296.

6. Discussion

This study showed that in terms of percent correct, forgetting was slower at first and then increased in the second interval. This finding contrasts with most previous research in which forgetting is relatively quick at first and appears to slow down (e.g., Ebbinghaus, 1885; Loftus, 1985).

The distribution model predicts this finding for the reasons illustrated in Figures 2–4. It seems safe to assume that individual items were forgotten in accordance with something like a power or exponential function (i.e., quickly at first and then more slowly). Because of the high rate of initial recall and (we assume) small standard deviation of item strengths, the distribution model predicts that the effect of the power or exponential function was weaker than the opposing effect of the distribution. This finding is a more extreme version of the apparent linear forgetting that has been found in other studies (Fisher & Radvansky, 2019); in this case, the distribution caused the curve to go past linear and become concave.

It is worth noting that the retention intervals in this study were relatively short compared to other forgetting research (e.g., Wickens, 1998; Wixted, 2004). Whether these findings would generalize to longer time scales is a question for future research. Relatedly, the short retention intervals led to high levels of performance. Thus, we cannot rule out the possibility that ceiling effects played a role in these findings. However, it is not clear that ceiling effects would produce a concave forgetting curve, and moreover, very high levels of performance may be necessary to observe concave forgetting, according to the distribution model.

7. A More Complex Distribution Model

Thus far, we have considered a relatively simple distribution model with two key assumptions: Items have a distribution of different memory strengths, and there is a recall threshold. A more complex model could make additional assumptions; for example, the position of items might not be the only thing that changes; if threshold positions were to change, it would also impact the shape of the empirical forgetting curve. Furthermore, item retrievability might momentarily fluctuate, and upward fluctuation could depend on what retrieval cues were present at the time of retrieval. Upward fluctuations in item strength or downward movement of the thresholds could both help explain the finding, from the reminiscence and hypermnesia literatures, that retrieval success can increase over time (Payne, 1987; Roediger & Payne, 1982).

We will examine what happens when a different assumption is added to the model, that different items have different forgetting rates. That is, two memories could differ in terms of both their current level of strength and the rate at which they lose strength (i.e., are forgotten). Our simple model assumed that each memory had a unique level of strength, but to capture both current strength and forgetting rate, a more complex model would need to assume that memories are characterized by more than one strength variable.

The New Theory of Disuse (Bjork & Bjork, 1992, 2006) assumes that a given memory is characterized by two kinds of strength. Retrieval strength is how accessible a memory is at the moment of retrieval, while storage strength is how strongly the memory is stored on a long-term basis. Retrieval strength is the only kind of strength that can be measured directly, and therefore it is the type of strength we have discussed thus far—it determines whether a memory is retrieved at any given time. Storage strength, by contrast, controls the rate at which the memory is forgotten; memories with more storage strength are forgotten at lower rates. Thus, the New Theory of Disuse provides a theoretical foundation for assuming current strength and forgetting rate are separate attributes of a memory.

A more complex version of the distribution model could assume that different items have different initial levels of strength and different forgetting rates—that is, that a given memory could be represented in two different distributions: a retrievability distribution and a separate forgetting rate distribution. Figure 6 illustrates such a bivariate distribution.

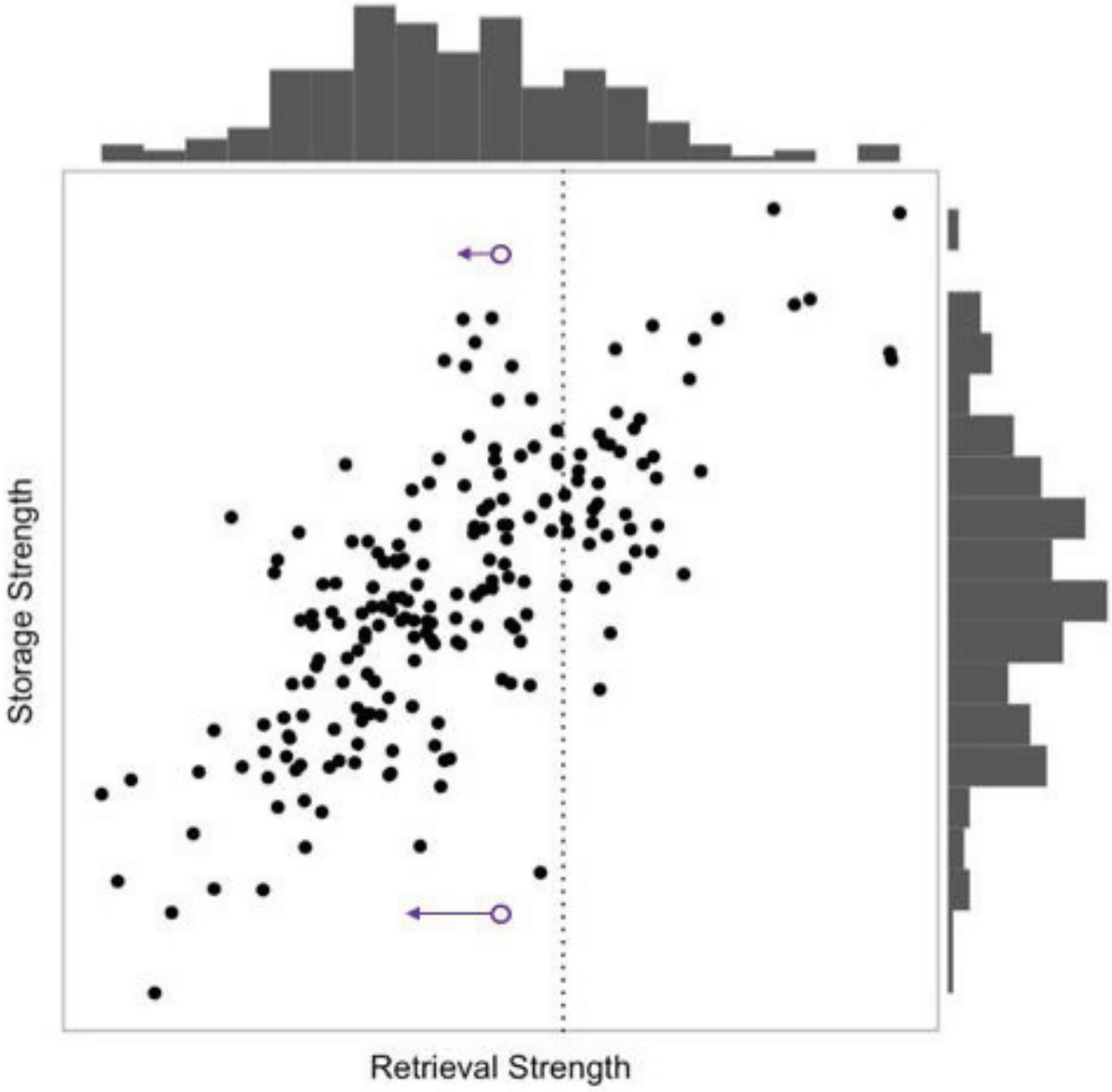


Figure 6. A hypothetical scatter plot with histograms in the margins. Each point in the scatter plot represents a memory. Retrieval strength is on the horizontal axis. The histogram above the scatter plot is the distribution of retrieval strengths based on the values in the scatter plot, which is comparable to the distribution of memory strengths in previous figures (e.g., Figure 1). The dashed vertical line represents the retrieval threshold. Storage strength is on the vertical axis, and it is represented by the histogram to the right of the scatter plot. Items with more storage strength tend to have more retrieval strength, creating a positive correlation. The two unfilled purple circles represent two items with the same initial retrieval strength. The lines attached to the circles represent the amount of forgetting in a set amount of time. The different line lengths show that items with less storage strength are forgotten (i.e., lose retrieval strength) more quickly than items with more storage strength.

Introducing variable forgetting rates into the model has the potential to make it significantly more complex. From a mathematical perspective, initial strength is a scalar value. It is easy to compute averages of scalar values. Forgetting rates, however, are represented by functions. If each item has its own forgetting function, then averaging across items becomes complicated because averaging across multiple functions does not necessarily return the same function. For example, when multiple exponential forgetting rates are averaged, the result is often fit best by a power function (R. B. Anderson, 2001;

R. B. Anderson & Tweney, 1997; Fisher & Radvansky, 2019; Murre & Chessa, 2011). This finding implies that when an empirical forgetting curve appears to be a power function, the underlying forgetting rates of the individual items might not be a power function.

Next, we describe a distribution model that includes variable forgetting rates. Such a model would depend on the nature of the forgetting function. The following equation is an exponential function:

$$f(t) = s - a^t$$

The variable t represents time, a is a constant that represents the item's forgetting rate, and s is the item's initial strength. In the simple distribution model described earlier, any given memory had its own unique value of s. In a more complex version of the model, each item would also have a unique value for a. Thus, instead of there being one distribution, there would be two, for s and a (see Figure 6).

A power function is similar to an exponential, but it has the following form:

$$f(t) = s - a \times t^b$$

A complex model could posit that there are three unique values for each item: s, a, and b. Thus, there could be three distributions, for s, a, and b.

If items are forgotten at different rates, then there is an additional complication: Over time there will be changes in the shape of the distribution of item retrieval strengths. Thus far, we have assumed that the mean, standard deviation, and shape of the distribution remain the same as they move leftward over time. However, differing forgetting rates should cause the shape of a distribution to morph as time passes. The nature of this morphing process depends upon the items' forgetting functions. It also depends on the correlation between an item's retrieval strength and its forgetting function; as Figure 6 shows, there should be a correlation between retrieval strength and storage strength—that is, between and item's current retrieval strength and its rate of forgetting—such that an item that is higher in retrieval strength will also tend to be forgotten more slowly. Because of this correlation, items with high retrieval strength (on the right side of Figure 6) also tend to have relatively high storage strengths, and thus they will move leftward (i.e., be forgotten) at relatively low rates; by contrast, items with lower retrieval strength will lose retrieval strength relatively quickly (i.e., items on the left side of Figure 6 will move leftward relatively quickly).

This morphing process has two implications. First, the distribution of retrieval strengths should grow wider and more negatively skewed over time. Second, above we made the assumption that the underlying distribution of memory strengths was bell-shaped (i.e., a normal distribution). This assumption was useful for illustration purposes, but the reality is that we do not know how memory strengths are initially distributed. Because of the morphing process, how they are distributed after a delay has no singular answer but rather changes over time.

We created a simple Monte Carlo model to simulate forgetting in a complex version of the distribution model. The key parameters of the model, which seem faithful to the underlying cognitive processes, were a mean and standard deviation for initial memory strength and a mean and standard deviation for each parameter in the forgetting functions. It was easy to closely fit empirical data using this model by playing around with parameter values. Power functions and exponential functions could fit the same dataset equally well, and the data could be fit using different assumptions about the shape of the initial memory strength distribution (e.g., normal or rectangular). We mention this model for two main reasons. The first is to illustrate the key parameters. The second is to point out a weakness: When a model has so many free parameters, its ability to fit data is not good evidence

that it is an accurate model (Roberts & Pashler, 2000). Furthermore, it is difficult to test such a model because it can be used to predict so many patterns of data. Perhaps future refinements of this kind of modelling will be able to address these issues.

8. The Importance of the Threshold

It is worth noting that the threshold is a crucial aspect of the distribution model because all the predictions of the model stem from the position of items relative to the threshold. The idea of a threshold goes hand in hand with the fact that correct answers are either recalled or not recalled (i.e., above or below the threshold).

When responses are either correct or incorrect, it becomes impossible to directly observe an item's forgetting rate—one can only observe whether it is currently above or below the threshold. If memory strength could be measured in a way that did not involve a threshold, it might be possible to observe individual forgetting rates unaffected by the shape or position of the distribution. One possible measure is response time, which is a continuum without a threshold. Response time is typically analyzed only after correct responses, which neutralizes the effect of the threshold because all responses would be for items that were above the threshold. Future research could use response time-based paradigms to shed light on item forgetting rates.

9. Educational Implications

Forgetting is an important aspect of education because there is little point in learning if the memories do not stick (P. C. Brown et al., 2014). The distribution model of forgetting has implications for education. Students often prepare for exams by studying a fact or concept enough to get the answer right on the exam, but not more. Doing this is similar to learning to criterion, where the criterion is the minimum amount of study required for good performance. Essentially, they reach a certain level of confidence and, to save time, decide "I can stop studying this one." This strategy is understandable and extremely common.

The distribution model suggests that this strategy has an important downside: Studying just enough to ace an exam will save time and produce good grades, but it will also maximize the loss of knowledge due to forgetting. To understand why, consider a student who studies for a quiz with 20 questions. Studying just enough means pushing those 20 items just slightly above their recall threshold, but not further. When a memory is near a threshold, a short period of time is enough for forgetting to drop it below the threshold. Thus, the distribution model can help explain why students who do well on exams frequently find that they do not remember what they learned for very long: Regardless of the shape of the forgetting curve, items close to the threshold will soon be forgotten.

Similarly, for educators, the distribution model helps explain an all-too-common outcome: When students do well on an exam, it is not safe to assume they have achieved long-term learning. Studying beyond a minimum threshold takes effort, but it also creates long-lasting memories, which is a core goal of education.

10. Concluding Comment

We have argued that distributions play a crucial, though often overlooked, role in the shape of empirical forgetting curves. In future research, considering and accounting for the impact of distributions might be a step toward understanding the underlying item forgetting rates, which are arguably the key cognitive process in forgetting.

Considering distributions might also be useful in other areas of psychology research. For example, distributions of memory strength relative to thresholds are relevant when learning causes memory strength to increase. In psychology more generally, if thresholds

are relevant in an area of research, then it may be worthwhile to consider the movement of distributions relative to the thresholds.

Author Contributions: Conceptualization, N.K. and R.A.B.; methodology, N.K.; formal analysis, N.K.; writing—original draft, N.K.; writing—review and editing, N.K. and R.A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Williams College (protocol code #2019-16; date of approval 21 March 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Anderson, M. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*(4), 415–445. [CrossRef]

Anderson, R. B. (2001). The power law as an emergent property. *Memory and Cognition*, *29*(7), 1061–1068. [CrossRef]

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, *25*(5), 724–730. [CrossRef]

Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, *55*(1), 25–35. [CrossRef]

Bahr, S. (2023, June 4). Why we write explainers: An explainer. *The New York Times*. Available online: https://www.nytimes.com/2023/06/04/insider/why-we-write-explainers-an-explainer.html (accessed on 21 May 2025).

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healey, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Routledge.

Bjork, R. A., & Bjork, E. L. (2006). Optimizing treatment and instruction: Implications of a new theory of disuse. In L.-G. Nilsson, & N. Ohta (Eds.), *Memory and society: Psychological perspectives* (pp. 116–140). Psychology Press.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*(1), 12–21. [CrossRef]

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Harvard University Press.

Donkin, C., & Nosofsky, R. M. (2012). A power-law model of psychological memory strength in short- and long-term recognition. *Psychological Science*, *23*(6), 625–634. [CrossRef]

Ebbinghaus, H. E. (1885). *Memory: A contribution to experimental psychology*. Henry A. Dover.

Fisher, J. S., & Radvansky, G. A. (2019). Linear forgetting. *Journal of Memory and Language*, *108*, 104035. [CrossRef]

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505–513. [CrossRef]

Hardt, O., Nader, K., & Nadel, L. (2013). Decay happens: The role of active forgetting in memory. *Trends in Cognitive Sciences*, *17*(3), 111–120. [CrossRef]

Knight, J. B., Ball, H. B., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language*, *66*(4), 731–746. [CrossRef]

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. [CrossRef]

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–998. [CrossRef] [PubMed]

Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 397–406. [CrossRef]

Mensink, G.-J., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review*, *95*(4), 434–455. [CrossRef]

Murre, J. M. J., & Chessa, A. G. (2011). Power laws from individual differences in learning and forgetting: Mathematical analyses. *Psychonomic Bulletin and Review*, *18*(3), 592–597. [CrossRef] [PubMed]

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review*, *115*(3), 544–576. [CrossRef]

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, *76*, 1–25.

Payne, D. G. (1987). Hypermnesia and reminiscence in recall: A historical and empirical review. *Psychological Bulletin*, *101*(1), 5–27. [CrossRef]

Rivera-Lares, K., della Sala, S., Baddeley, A., & Logie, R. (2023). Rate of forgetting is independent from initial degree of learning across different age groups. *Quarterly Journal of Experimental Psychology*, *76*(7), 1672–1682. [CrossRef]

Rivera-Lares, K., Logie, R., Baddeley, A., & della Sala, S. (2022). Rate of forgetting is independent of initial degree of learning. *Memory & Cognition*, *50*(8), 1706–1718. [CrossRef]

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367. [CrossRef]

Roediger, H. L., & Payne, D. G. (1982). Hypermnesia: The role of repeated testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*(1), 66–72. [CrossRef]

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. [CrossRef] [PubMed]

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*(1), 49–60. [CrossRef] [PubMed]

Wickens, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel (1996): A quantitative description of retention. *Psychological Review*, *105*(2), 379–386. [CrossRef]

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*(1), 235–269. [CrossRef]

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science*, *2*(6), 409–415. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Exploring Whether Making Second-Language Vocabulary Learning Difficult Enhances Retention and Transfer

Alice F. Healy [1,*], Vivian I. Schneider [1] and James A. Kole [2]

[1] Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, CO 80309-0345, USA; vicki.schneider@colorado.edu
[2] School of Psychological Sciences, University of Northern Colorado, Greeley, CO 80639, USA; james.kole@unco.edu
* Correspondence: alice.healy@colorado.edu; Tel.: +1-303-579-0576

Abstract

Four previous and two new experiments from our laboratory are reported, in which college students learned associations between French and English words in a learning phase and then took an immediate retention test. One week later, a delayed test was followed by relearning. Four difficulty manipulations were used during learning: blocking versus mixing semantic categories, translation direction, prelearning, and set size. The first new experiment examined the effect of set size on retention, and the second new experiment examined blocking and mixing semantic categories, as well as translation direction, on learning new vocabulary (transfer). Generally, across the six experiments, difficult conditions provided a disadvantage during learning and immediate testing, but made no difference or provided an advantage during relearning and delayed testing. These results suggest that making the initial learning more difficult does not always lead to superior retention.

Keywords: vocabulary learning; retention; transfer; desirable difficulties

1. Introduction

There are many advantages to possessing a large vocabulary. Studies have shown that knowledge of words' meanings is crucial for effective communication, enhancing both verbal (e.g., Jongman et al., 2021; Shao et al., 2014; Unsworth et al., 2010) and written (e.g., Olinghouse & Leaird, 2009; Olinghouse & Wilson, 2013) expression. Other studies have shown that vocabulary knowledge is strongly related to reading speed and comprehension (Currie & Muijselaar, 2019; Wagner & Meros, 2010). Moreover, a larger vocabulary contributes to cognitive development, promoting critical thinking and problem-solving abilities, and allowing individuals to engage in more complex reasoning (Kievit et al., 2017, 2019).

Similarly, learning a *second* language also offers numerous cognitive benefits that extend beyond expression and reading comprehension. Bilingualism has been linked to enhanced cognitive flexibility, improved memory, and better multitasking abilities, making bilingual individuals more adept at executive control tasks (Bialystok, 2011; but see Paap et al., 2015, for a different conclusion). Additionally, research suggests that bilingualism may provide neuroprotective effects, potentially delaying the age of onset for symptoms of Alzheimer's disease (Bialystok, 2021).

Given its importance, many studies have explored methods to improve second language learning, with a strong focus on contextual interference. Contextual interference

occurs when multiple skills or variations of a task are interleaved during learning, rather than being practiced in a blocked format (see Battig, 1972, 1979, for the development of the concept of contextual interference). Although this approach leads to slower learning rates, it ultimately enhances both retention and transfer (for a review, see Czyż et al., 2024).

An important early study demonstrating the contextual interference effect was conducted by Shea and Morgan (1979) in the domain of motor skill learning. Subjects learned three distinct movement patterns, each requiring them to knock down three barriers in a specific sequence using their hand. During training, one group practiced in a blocked format, completing multiple trials of one pattern before switching to the next, while another group practiced in a random or interleaved format, alternating between patterns on each trial. Subjects were tested twice: 10 min after training and again 10 days later. The tests included the previously practiced movement patterns (retention), as well as two novel patterns that had not been introduced during training (transfer). The results showed that blocking had different effects on training and test performance. During training, the blocked group performed the movements more quickly than the random group; however, the random group performed better on both the studied patterns and novel patterns during testing, thereby demonstrating better retention and transfer. More recent research has confirmed these findings and expanded on the role of contextual interference in skill acquisition (e.g., Mousavi et al., 2024).

Subsequent research has demonstrated that contextual interference is a more general principle that extends beyond motor learning. For example, in visual category learning tasks for which subjects are shown exemplars of different categories, training might be blocked, such that subjects are presented with exemplars from only one category at a time, or training might be interleaved, such that the exemplars from different categories are intermixed. In general, contextual interference leads to increased retention of the studied exemplars, as well as increased transfer, or the ability to categorize new exemplars (e.g., Carvalho & Goldstone, 2014; Kang & Pashler, 2012; Kornell & Bjork, 2008). Contextual interference has also been shown to improve retention and transfer in more educationally relevant materials, such as chemistry concepts (Eglington & Kang, 2017), and tasks such as classifying statistics problems (Sana et al., 2018) and mathematical problem solving (Mielicki & Wiley, 2022; Rohrer & Taylor, 2007). However, the extent to which contextual interference benefits retention and transfer might vary by task type, with some tasks showing stronger benefits from interleaving than others (Brunmair & Richter, 2019). In certain cases, blocking might enhance retention more than interleaving would (e.g., Carpenter & Mueller, 2013; Sorensen & Woltz, 2016).

Schneider et al. (1998) investigated contextual interference as it relates to second-language vocabulary learning. In two experiments, subjects studied French–English word pairs from five different semantic categories (e.g., body parts or articles of clothing). The study compared two learning conditions: a blocked condition, in which each block of trials during the learning phase included words from only one semantic category, and a mixed condition, in which each block of trials during the learning phase included one word from each semantic category. Subjects in Experiment 1 underwent learning rounds until they responded correctly to all word pairs two times in a row, whereas subjects in Experiment 2 all underwent exactly three learning rounds. Following learning, subjects took an immediate retention test as well as a delayed retention test one week later. Following the delayed retention test, subjects completed three more learning rounds to assess relearning. For both experiments, subjects in the blocked condition demonstrated better learning, but worse relearning after the delay. These findings suggest that although blocking by semantic category might facilitate initial learning, it might not promote long-term retention as effectively as mixing or interleaving semantic categories. However, research on the

effects of mixing versus blocking semantic categories during second language vocabulary acquisition has yielded mixed results. Some studies have found that blocking words into semantic categories hinders learning (e.g., Bolger & Zapata, 2011; Tinkham, 1993, 1997), possibly due to the physical similarity between the referents of the words (Bach & Barclay, 2025). In contrast, other studies reported the opposite effect, with blocked presentation leading to better learning outcomes (Hashemi & Gowdasiaei, 2005; Hoshino, 2010).

Contextual interference can be conceptualized as one of a broader class of learning strategies known as *desirable difficulties* (R. A. Bjork, 1994; R. A. Bjork & Bjork, 2020; R. A. Bjork & Kroll, 2015). Desirable difficulties are conditions that introduce challenges during learning, which might initially lead to lower accuracy or a slower learning rate but might ultimately enhance long-term retention and transfer (E. L. Bjork & Bjork, 2011). These difficulties can take many forms, such as the *testing effect* (e.g., E. L. Bjork et al., 2014; for a review, see Yang et al., 2021), in which retrieval practice via testing strengthens memory more than passive restudy, and the *generation effect* (McNamara & Healy, 2000; for a review, see McCurdy et al., 2020), where producing information rather than simply reading it improves recall. Other examples include the *spacing effect* (e.g., Carpenter, 2017), in which learning is distributed over time rather than being massed in a single session, and *variability of practice*, which exposes learners to diverse contexts or problem types (e.g., Raviv et al., 2022).

In a follow-up to their 1998 study, Schneider et al. (2002, Experiment 1) investigated whether contextual interference, as well as the potential desirable difficulty of a hard translation direction, enhances both the retention and transfer of second language vocabulary items. Second language vocabulary learning is inherently bi-directional, requiring two complementary translation processes: one from the second language to the first and the other in the reverse direction. Experiment 1 examined both translation processes and the transfer between them. Specifically, half of the subjects (all of whom did not speak French) completed a learning phase and an immediate retention test using French words as cues and English words as responses, representing a less difficult translation task. The remaining subjects completed a learning phase and an immediate retention test with English words as cues and French words as responses, a more challenging translation task. All subjects completed three learning rounds before the immediate retention test.

In a second session, conducted one week after learning, subjects completed a delayed retention test followed by relearning. During relearning, they either continued with the same translation direction as in the learning phase or switched to the opposite direction. To examine all possible combinations of first and second tasks, four groups were employed: (a) French-to-English first, French-to-English second; (b) English-to-French first, English-to-French second; (c) French-to-English first, English-to-French second; and (d) English-to-French first, French-to-English second. A contextual interference manipulation was also included, as in Schneider et al.'s (1998) study, with items either blocked by semantic category or mixed. Thus, each group was further divided into four subgroups based on the practice schedules used across the two sessions: (e) blocked first, blocked second; (f) mixed first, mixed second; (g) blocked first, mixed second; and (h) mixed first, blocked second.

As predicted, the manipulation of translation direction had the opposite effects on learning and relearning, similar to the earlier demonstrations of contextual interference. Specifically, accuracy was much higher during learning for the French-to-English translation direction than for the English-to-French translation direction. This advantage was also evident in the immediate test. However, subjects given the more difficult English-to-French translation direction during learning showed higher accuracy on the delayed retention test one week later, as well as somewhat higher accuracy during relearning, regardless of the translation direction used during relearning. The latter result demonstrates that increasing

the difficulty of the task during learning aids in the long-term retention of the original task, as well as in transfer to a new task.

Other studies obtained similar results regarding translation direction and retention. Griffin and Harley (1996) found that presenting first-language cues and requiring second-language responses (L1-L2) during learning, as in the English-to-French translation direction of Schneider et al. (2002), resulted in greater retention on comprehension and production tests than the easier L2-L1 (second language cues, first language responses) translation direction, possibly due to the increased retrieval demands and deeper processing. Others, including Terai et al. (2021), revealed a similar pattern, but only when subjects were more proficient with the second language. Webb (2009) found that L1-L2 translation direction increased retention as well as orthographic and syntactic knowledge.

Schneider et al. (2002, Experiment 2) introduced a second manipulation focused on learning difficulty (in this case *reducing* it) by providing prelearning for half of the French words. For prelearning, subjects were presented with a French word and simply had to type it. Each prelearned word appeared three times across three prelearning rounds. Following prelearning, as in Experiment 1, subjects completed a learning phase and an immediate retention test, as well as a delayed retention test and a relearning phase during a second session one week later. Also, as in Experiment 1, half of the subjects were assigned the more difficult English-to-French translation direction, whereas the other half were assigned the less difficult French-to-English direction. In the second session, half of the subjects in each condition continued with the same translation direction, whereas the other half switched to the opposite direction for both the delayed retention test and relearning. For Experiment 2 there was not a blocked/mixed manipulation; for all subjects, the words were mixed across semantic categories.

As in Experiment 1, subjects in the French-to-English translation condition performed better during learning than those in the English-to-French translation condition. However, during relearning in Session 2, those in the English-to-French condition showed greater improvement across sessions, despite initially performing worse. Similarly, subjects assigned to the easier French-to-English condition during learning outperformed those in the English-to-French condition on the immediate test, but this advantage slightly reversed on the delayed retention test.

A comparable pattern emerged for prelearned versus non-prelearned words. During learning in Session 1, subjects performed better on prelearned words than on non-prelearned words. However, by Session 2, the performance improvements were greater for non-prelearned words than for prelearned words during relearning. Despite these differences in learning, prelearning had no effect on test performance.

Another difficulty manipulation explored in both first- and second-language vocabulary learning is block size, or the number of words studied together before testing. In a study of first-language vocabulary, Kornell (2009) found that studying all vocabulary items in a single block led to better retention than dividing them into four smaller blocks. Nakata and Webb (2016) investigated block size in second-language learning and found that larger blocks led to a reduced performance during learning rounds. However, block size only had a consistent effect on final test performance when controlling for spacing, or the interval between the repeated presentations of a word or block. Similarly, Nakata et al. (2023) reported that increasing spacing during learning enhances retention, and that this benefit becomes even more pronounced when spacing is repeated during relearning phases.

In summary, contextual interference occurs when learning tasks are interleaved rather than blocked. As demonstrated in early studies of motor learning (Shea & Morgan, 1979), this effect was subsequently observed in category learning (e.g., Carvalho & Goldstone, 2014; Kang & Pashler, 2012; Kornell & Bjork, 2008), chemistry (Eglington & Kang, 2017),

mathematics (Mielicki & Wiley, 2022; Rohrer & Taylor, 2007), statistical problem-solving (Sana et al., 2018), and the acquisition and retention of logic rules (Schneider et al., 1995). Contextual interference is just one of many desirable difficulties that lead to slower initial learning but enhance retention and transfer. In previous studies by (two of) the present authors (Schneider et al., 1998, 2002), both contextual interference and desirable difficulties were examined in a second-language vocabulary learning task. These studies showed that increasing difficulty, whether through contextual interference, translation direction, or a prelearning manipulation, results in better performance during learning, but reduced performance in a delayed test or during relearning.

The present study follows Schneider et al. (1998, 2002) by providing two new experiments that examine other manipulations of task difficulty and their effects on learning, relearning, retention, and transfer on a task involving learning a second language.

Unlike earlier studies by others, using varied methodologies and often yielding contradictory results, we use the same general methodology in these new experiments as in our earlier studies, enabling us to examine vocabulary learning, relearning, retention, and transfer in a consistent, controlled way across different task-difficulty manipulations.

2. Experiment 1

In Experiment 1, to manipulate difficulty, we varied the block size, or the number of words that were presented together during learning. Previous studies have also examined this issue. Kornell (2009) compared block sizes of 4 and 20, whereas Nakata and Webb (2016) compared block sizes of 4, 10, and 20. In the present experiment, subjects either completed the learning phase with a smaller set size, where each block included only six words, or with a larger set size, where each block included twelve words. Experiment 1 focused on retention rather than transfer, with subjects completing an immediate test and a delayed test (as in Schneider et al., 1998, 2002), as well as a relearning phase. Increasing the block set size might create a desirable difficulty, such that retention is enhanced. However, increasing difficulty might also result in cognitive overload, which impairs retention and transfer (Xie et al., 2017).

2.1. Method

2.1.1. Subjects

A total of 48 University of Colorado undergraduates participated as subjects for credit in an introductory psychology course. All of these subjects were native English speakers who did not speak French. Previous studies examining block size and retention (e.g., Kornell, 2009; Nakata & Webb, 2016), typically yielded large effect sizes. A power analysis based on Kornell (2009; $\eta^2 = 0.21$) suggests that 32 subjects would be sufficient to achieve 0.8 power for the main effect of set size in the present experiment. An analysis based on Nakata and Webb (2016, Experiment 2; $\eta^2 = 0.41$) indicates that as few as 14 subjects are needed to achieve 0.8 power. The difficulty manipulation (prelearning) used by Schneider et al. (2002; $\eta^2 = 0.25$) suggests that 26 subjects would be adequate to achieve sufficient power. These studies varied in terms of factors such as whether they examined first- or second-language vocabulary learning, the difficulty manipulation used, the language learned, and the set size.

2.1.2. Materials

Thirty-six French words were used in this experiment, including six words from each of six different semantic categories: body parts, school, transportation, dining, food, and clothing. (For a list of the 36 items used in this study, see Appendix A).

2.1.3. Procedure

Experiment 1 included two sessions that were spaced one week apart. The first session included a learning phase and an immediate retention test; the second session included a delayed retention test followed by a relearning phase.

For both the learning and relearning phases, in each trial, subjects were presented with a French word and its English translation for a period of 2 s. In the *small set size* condition, one word was presented from each of six categories, and after six French word–English translation pairs were presented, subjects were tested over the preceding six items. Subjects were provided with the French word and had to type the corresponding English translation. This learning/test cycle repeated until the 36 words were presented and tested three times. The *large set size* condition proceeded in a similar manner. However, subjects in this condition were tested after 12 French word–English translation pairs were presented, rather than 6, and the 12 pairs included two words from each of six categories. During learning, half of the subjects were assigned to the small set size condition and the other half to the large set size condition; during relearning, half of the subjects remained in the same condition they were in during learning, whereas the other half switched to the other set size condition.

For both the immediate and delayed retention tests, subjects were tested with all 36 words via cued recall, as occurred during the learning rounds. More specifically, they were tested in the French-to-English direction over all 36 items without further study.

2.1.4. Design

For the analyses of learning phases, the design was a $2 \times 2 \times 2 \times 3$ mixed factorial. There were two between-subjects variables: set size for the learning phase (large, small) and set size during relearning. There were also two within-subjects variables: session (Session 1 and Session 2) and learning round (Round 1, Round 2, and Round 3).

For the analysis of the test, the design was a simpler $2 \times 2 \times 2$ mixed factorial including the variables of set size during learning (large, small), set size during relearning (large, small), and test (immediate, delayed).

The dependent variable for both learning and testing was the proportion of correct responses.

2.2. Results

For both Experiments 1 and 2, we analyzed our results using analyses of variance (ANOVAs).

2.2.1. Learning

For learning, there was a main effect of session ($F(1, 44) = 236.720$, $MSE = 0.014$, $p < 0.0001$, and $\eta^2 = 0.843$), reflecting the learning from the first ($M = 0.675$) to the second ($M = 0.891$) session. Learning was also reflected in the significant main effect of learning round ($F(2, 88) = 257.418$, $MSE = 0.004$, $p < 0.0001$, and $\eta^2 = 0.854$), with an increase from the first round ($M = 0.673$) to the second round ($M = 0.809$), and to the third round ($M = 0.868$). There was also a significant interaction between session and learning round ($F(2, 88) = 107.872$, $MSE = 0.003$, $p < 0.0001$, and $\eta^2 = 0.710$), reflecting the greater improvement across rounds observed in the first session compared to the second (see Figure 1). Most importantly, there was also an interaction between learning set size and session, ($F(1, 44) = 9.639$, $MSE = 0.014$, $p < 0.0001$), and $\eta^2 = 0.180$. Specifically, overall accuracy was higher for subjects in the small learning set size condition than for those in the large learning set size condition in Session 1, but was virtually the same in Session 2 (see Figure 2). Thus, the improvement across sessions was greater for the large learning set size condition than for small learning set size condition.

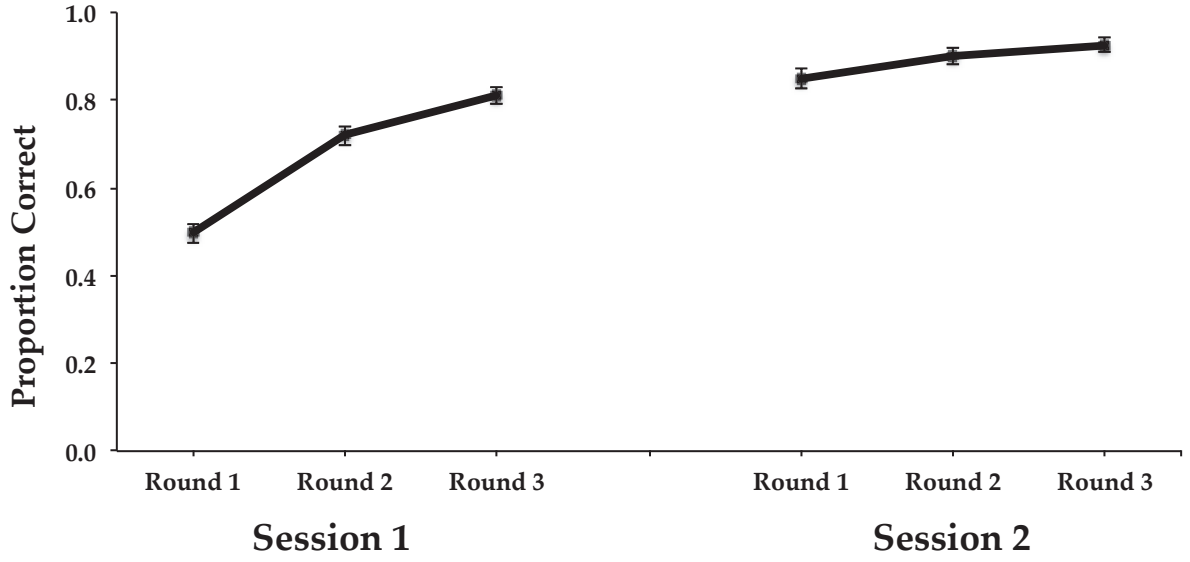


Figure 1. Proportion of correct answers during learning and relearning in Experiment 1 as a function of learning round (Round 1, Round 2, Round 3) and session (Session 1, Session 2). Error bars represent between-subjects standard errors of the mean.

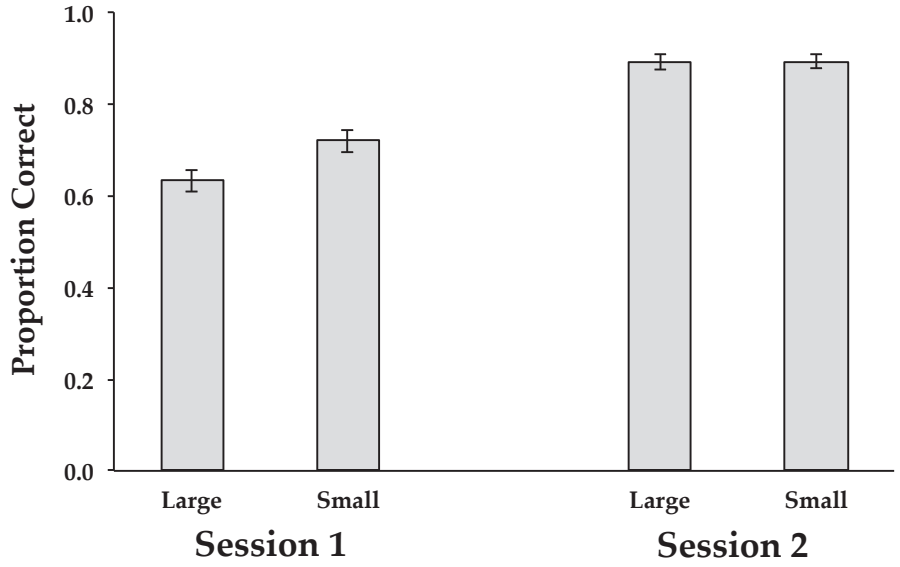


Figure 2. Proportion of correct answers during learning and relearning in Experiment 1 as a function of learning set size (Large, Small) and session (Session 1, Session 2). Error bars represent between-subjects standard errors of the mean.

2.2.2. Test

The only significant effect during testing was the main effect of session, $F(1, 44) = 165.150$, $MSE = 0.005$, $p < 0.0001$, $\eta^2 = 0.790$, reflecting words that were forgotten between the immediate test ($M = 0.727$) and the delayed test ($M = 0.541$).

2.3. Summary and Discussion

Experiment 1 manipulated difficulty by varying the set size or the number of items that were presented and tested at a time. As with other studies investigating desirable difficulties, increasing difficulty had different effects on learning and retention over time. Initially, those in the small learning set size condition performed better, but by the second session, accuracy was similar across learning set size conditions. This finding suggests that

the large learning set size condition, despite being more difficult initially, did not impair relearning. This finding is also interesting given the research on procedural reinstatement, which states that retention and transfer are maximized when the learning and testing procedures match (Healy et al., 1992). According to procedural reinstatement, those who were in the same set size condition during learning and relearning should have performed better than those who switched. However, there was no advantage during relearning for those in the same condition as they were in during learning. There were also no differences between the learning set size conditions regarding the test results, which might suggest that there is no advantage in terms of retention of using larger set sizes, although we did not control for spacing (Nakata & Webb, 2016). However, as research on the testing effect has shown, successful retrieval during learning might not impact test performance (e.g., Chan et al., 2024).

3. Experiment 2

Experiment 2 explored the extent to which transfer is affected by increasing task difficulty. Until this point, our studies on transfer were limited to variations in the translation direction (Schneider et al., 2002). An examination of transfer from one list of word pairs to a completely different list is equally important. In Experiment 2, we investigated this type of transfer. Subjects learned words from six semantic categories during the first session; during the second session, subjects learned new words from the same six semantic categories, as well as new words from different semantic categories. The learning rates and memory for the two types of words were compared to determine whether previous experience with the semantic categories facilitates the learning of new words from those categories. Generally, research shows that prior knowledge can facilitate the learning of domain-related information (e.g., Kole & Healy, 2007; Van Overschelde & Healy, 2001), whereas specific studies of second-language vocabulary learning show that blocking words into semantic categories can hinder learning (e.g., Bolger & Zapata, 2011; Tinkham, 1993, 1997). However, the present study focuses on transfer to the learning of new items.

3.1. Method

3.1.1. Subjects

Again we employed a sample of 48 University of Colorado undergraduate students participating as subjects for credit in an introductory psychology course. All of these subjects were native English speakers who did not speak French. Power estimates were based on studies by Tinkham (1993, 1997) because they were methodologically similar to the present experiment. Specifically, vocabulary items were presented as word pairs, and the primary dependent variable was a measure of accuracy. For these studies, the η^2 for the main effect of relatedness (grouping semantic items vs. unrelated items) was large (Tinkham, 1993, $\eta^2 = 0.59$; Tinkham, 1997, Experiment 1, $\eta^2 = 0.51$, and Experiment 2, $\eta^2 = 0.37$). These effect sizes suggest that between 8 and 16 subjects would be sufficient to achieve 0.8 power. However, the focus of the present study is on whether or not semantic categories influence the learning of new items rather than on how well the items in the semantic categories are learned relative to semantically unrelated items.

3.1.2. Materials

The materials were 96 French–English word pairs (12 words from each of eight semantic categories). (For a list of the 96 items used in this study, see Appendix B).

3.1.3. Procedure

For the first session, subjects completed the learning phase, where they learned 36 word pairs; the words were presented in a mixed format, with one word from each of the six semantic categories in each block. As in Experiment 1, each French word/English translation was presented for 2 s, and after six French words/English translations were presented, subjects were tested using cued recall. Subjects completed three learning rounds, with each of the 36 word pairs being presented and tested three times. Subjects were given an immediate test of all 36 word pairs. For the purpose of analysis, these 36 word pairs were classified into three categories, SW, SC, and DC, based on the counterbalancing scheme. Across subjects, counterbalancing ensured that word pairs were assigned equally often to each of these categories.

For the second session, which was held one week later, subjects completed a delayed test, followed by three rounds of learning, and then a final test. Subjects again learned and were tested over a total of 36 words. However, only 12 words were the same as those in the learning phase. Those 12 words included 6 words from each of two of the six categories employed during Session 1 (SW; same words). Another 12 words were 6 new words from each of another two categories employed in Session 1 (SC; same category), and the third set of 12 words consisted of 6 new words from each of two new categories that were not previously employed during the learning phase (DC; different category). Thus, the breakdown of word pairs into categories was meaningful only in the second session.

In the first week, half of the subjects learned and were tested in a French-to-English translation direction, and the remaining half learned and were tested in an English-to-French translation direction. For the second week all subjects were tested and relearned in the harder English-to-French direction.

3.1.4. Design

The design for the analysis of learning performance was a $2 \times 2 \times 3 \times 3$ mixed factorial, with translation direction during learning (English-to-French or French-to-English), as the between-subjects variable, session (Session 1 or Session 2), learning round (Round 1, Round 2, or Round 3), and word type (SW words used in both sessions, SC words used only during relearning from previously studied categories, or DC words used only during relearning from unstudied categories) as the within-subject variables.

The design of the analysis of test performance was a $2 \times 3 \times 3$ mixed factorial, with translation direction during learning in Week 1 (English-to-French or French-to-English) as the between-subjects variable, and test (immediate, delayed, or final), and word type (SW, SC, or DC) as the within-subject variables.

The dependent measure during the learning rounds and the tests in each of the two sessions was the proportion of correct responses.

3.2. Results

3.2.1. Learning

Overall performance during learning in Session 1 was much better for subjects learning in the French-to-English direction than for those given the English-to-French direction; however, there was no difference between the two groups during relearning in Session 2 (see Figure 3). Thus, there was an interaction between translation direction during learning and session, $F(1, 46) = 117.369$, $MSE = 0.049$, $p < 0.0001$, and $\eta^2 = 0.718$, as well as a main effect of translation direction during learning, $F(1, 46) = 19.433$, $MSE = 0.295$, $p < 0.0001$, and $\eta^2 = 0.297$, and a main effect of session, $F(1, 46) = 4.218$, $MSE = 0.049$, $p = 0.0457$, and $\eta^2 = 0.084$.

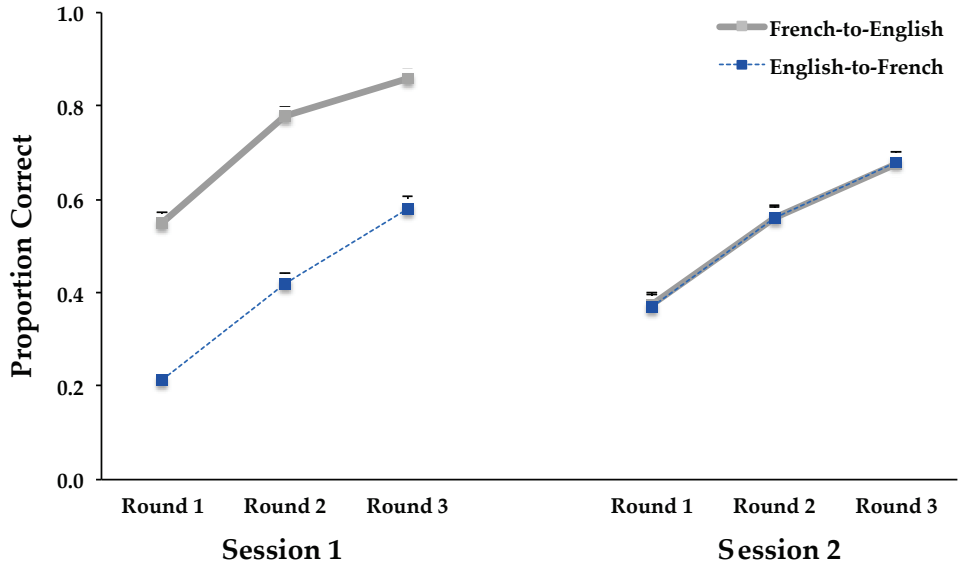


Figure 3. Proportion of correct answers during learning and relearning in Experiment 2 as a function of learning translation direction (French-to-English, English-to-French), session (Session 1, Session 2), and learning round (Round 1, Round 2, Round 3). Error bars represent the between-subjects standard errors of the mean.

Both the French-to-English and the English-to-French groups improved across learning rounds in both sessions; however, during relearning, when both groups performed the more difficult task of translating from English to French, the English-to-French group improved greatly from Session 1 whereas the French-to-English group declined in accuracy from Session 1 (see Figure 3). The interaction of translation direction, session, and learning round was significant ($F(2, 92) = 3.787$, $MSE = 0.007$, $p = 0.0263$, and $\eta^2 = 0.076$), as was the interaction of session and learning round ($F(2, 92) = 3.727$, $MSE = 0.007$, $p = 0.0278$, and $\eta^2 = 0.075$), and the main effect of learning round ($F(2, 92) = 396.290$, $MSE = 0.019$, $p < 0.0001$, and $\eta^2 = 0.896$).

A better performance was obtained during relearning in Session 2 for the SW word type, or words that were included in both learning and relearning phases, than for either set of other words only shown during relearning (SC, DC), with the advantage for the SW word type decreasing across relearning rounds (see Figure 4). The interaction between session, learning round, and word type was significant ($F(4, 184) = 4.960$, $MSE = 0.011$, $p = 0.0008$, and $\eta^2 = 0.097$), as was the main effect of word type ($F(2, 92) = 9.536$, $MSE = 0.065$, $p = 0.0002$, and $\eta^2 = 0.172$). The component interactions between word type and session ($F(2, 92) = 32.628$, $MSE = 0.024$, $p < 0.0001$, and $\eta^2 = 0.415$), and between word type and learning round ($F(4, 184) = 6.350$, $MSE = 0.010$, $p < 0.0001$, and $\eta^2 = 0.121$) were also significant.

To examine transfer, an analysis was conducted that was restricted to the three blocks of relearning during Session 2, and included only the SC and DC categories. Both SC and DC words were new during Session 2; however, subjects learned words from the same semantic categories as SC words during Session 1, but had no previous exposure to the semantic categories for DC words. This analysis revealed no significant main effects or interactions involving the word type variable. However, because subjects were tested over only six word pairs at a time during relearning blocks, the differences between these word types might have been minimized.

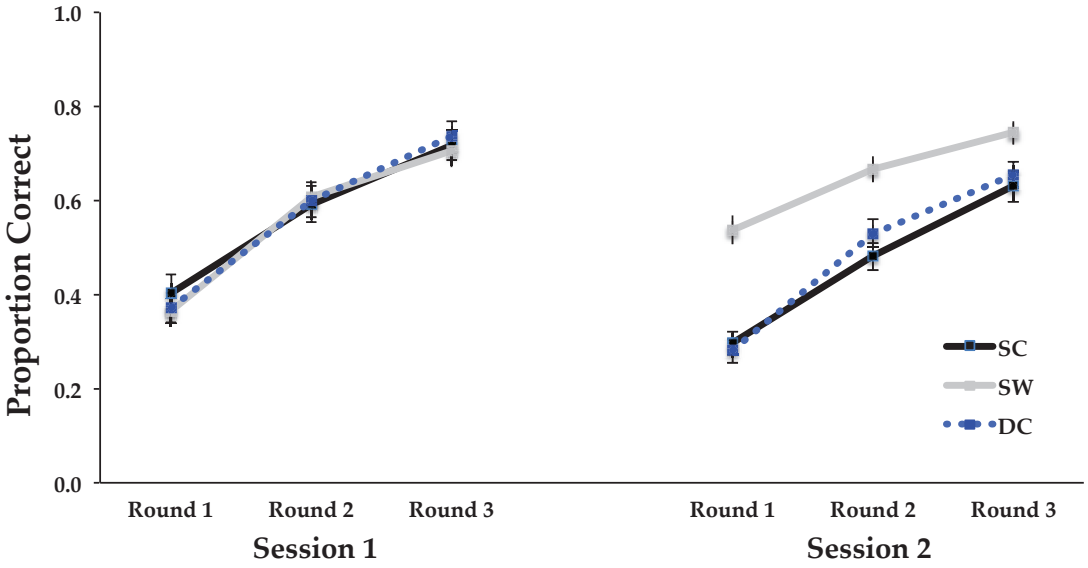


Figure 4. Proportion of correct answers during learning and relearning in Experiment 2 as a function of word type (different categories, DC; same words, SW; and same categories, SC), session (Session 1, Session 2), and learning round (Round 1, Round 2, Round 3). Error bars represent between-subjects standard errors of the mean.

3.2.2. Test

Overall, accuracy was higher for those who learned using the easier French-to-English direction than those who learned using the more difficult English-to-French direction; the main effect of translation direction during learning was significant: $F(1, 46) = 6.557$, $MSE = 0.124$, $p = 0.0138$, and $\eta^2 = 0.125$ (see Figure 5). Accuracy was also higher for the immediate and final tests than for the delayed test; the main effect of the test was significant: $F(2, 92) = 482.105$, $MSE = 0.032$, $p < 0.0001$, and $\eta^2 = 0.913$. The main effect of word type was significant as well, as, on average, across tests, accuracy was highest for SW words, lowest for SC words, and intermediate for DC words: $F(2, 92) = 16.415$, $MSE = 0.028$, $p < 0.0001$, and $\eta^2 = 0.263$.

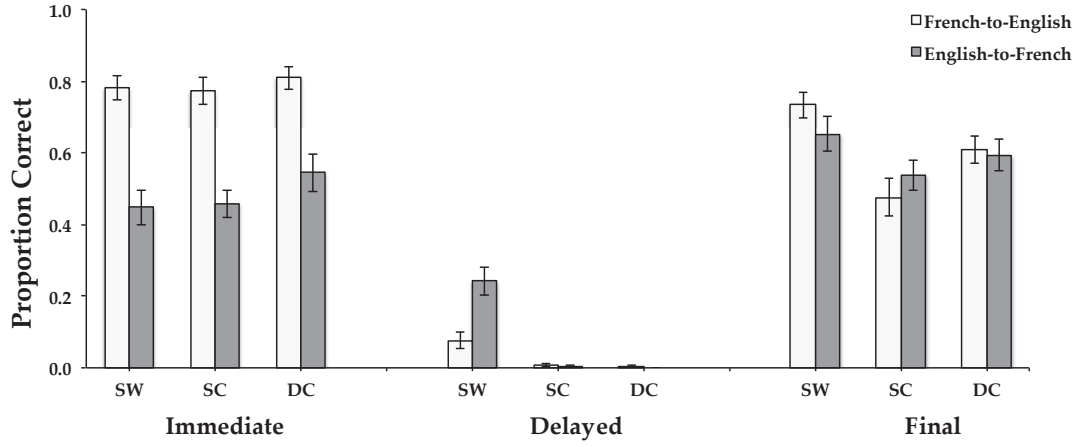


Figure 5. Proportion of correct test answers in Experiment 2 as a function of learning translation direction (French-to-English, English-to-French), test (Immediate, Delayed, Final), and word type (different categories, DC; same words, SW; and same categories, SC). Error bars represent between-subjects standard errors of the mean.

The interaction between translation direction, test, and word type was significant: $F(4, 184) = 7.329$, $MSE = 0.013$, $p < 0.0001$, and $\eta^2 = 0.137$. For the immediate test, there was a

more consistent advantage across word types for the French-to-English direction than for the English-to-French direction. However, for the delayed test, performance was better for the harder English-to-French direction than for the easier French-to-English direction for SW words, which were included during the learning session. Performance was worst for both types of new words (SC and DC, both of which included words that were not presented during the learning phase). For the final test, neither translation direction was consistently superior (see Figure 5). The component interactions between test and translation direction ($F(2, 92) = 41.227$, $MSE = 0.032$, $p < 0.0001$, and $\eta^2 = 0.473$), and between test and word type ($F(4, 184) = 15.483$, $MSE = 0.013$, $p < 0.0001$, and $\eta^2 = 0.252$), were also significant.

To test for transfer, a planned comparison between the SC and DC word types at the final test was conducted. The difference between word types was significant ($F(1, 46) = 13.587$, $MSE = 0.020$, $p = 0.0006$, and $\eta^2 = 0.228$) with a higher accuracy for the DC than for the SC word types (see Figure 5). Thus, negative transfer (or interference) occurred, as previous experience with semantic categories hindered memory of new words from those categories.

3.3. Summary and Discussion

Experiment 2 examined the effects of both translation direction difficulty and blocking words into semantic categories during second-language vocabulary learning. Previous studies have investigated blocking by semantic category (e.g., Bolger & Zapata, 2011; Tinkham, 1993, 1997), with the general finding that such blocking can impair retention. The present study also tested for transfer by examining learning and memory for new words belonging to previously studied categories, as well as unstudied categories.

Similarly to the study by Schneider et al. (2002; Experiment 1), the more difficult English-to-French translation direction resulted in lower accuracy during both the learning and the immediate test than the French-to-English translation direction; however, these differences were virtually eliminated during relearning. Thus, increasing the learning difficulty during the first session did not result in a superior performance during the second, although it did not decrease performance. Thus, at least according to the findings of this experiment, translation direction does not appear to be a desirable difficulty for second-language vocabulary learning.

Subjects demonstrated significant forgetting from the immediate test to the delayed test. However, there was some evidence of retention in that words that were presented during both sessions (SW) were learned more quickly than new words introduced during Session 2 (SC, DC). In terms of transfer, there was a clear *negative* transfer effect: during the final test, subjects recalled words from entirely new categories (DC) at a higher rate than new words from previously studied categories (SC). Although this result is somewhat surprising given previous research that knowledge of a domain makes learning domain-relevant information more efficient (e.g., Kole & Healy, 2007), second-language vocabulary learning studies consistently yielded the opposite pattern, presumably due to interference (e.g., Tinkham, 1993, 1997).

4. General Discussion

The present study investigated whether making second-language vocabulary learning more difficult enhances retention and transfer, extending the framework of desirable difficulties (R. A. Bjork, 1994; R. A. Bjork & Bjork, 2020; R. A. Bjork & Kroll, 2015) and contextual interference (e.g., Battig, 1972, 1979; Czyż et al., 2024). Across two experiments, we examined the effects of two difficulty manipulations, including set size during learning (Experiment 1), as well as translation direction and blocking according to semantic cat-

egory (Experiment 2), on second-language vocabulary acquisition, retention, relearning, and transfer.

Experiment 2 further explored transfer by assessing how the semantic relatedness between learning and relearning materials influences new learning. Previous studies (e.g., Tinkham, 1993, 1997) found that learning words in semantic clusters impairs the retention of those words compared to mixing words from different semantic categories. Experiment 2 examined transfer by assessing whether knowledge of exemplars from a semantic category negatively impacts the learning of new members of that category. Indeed, this pattern was observed in Experiment 2: subjects performed worse on words from previously studied categories (SC) than on words from entirely new categories (DC) during the final test, indicating a negative transfer effect. This result suggests that prior exposure to semantically related content can produce interference rather than facilitation, consistent with earlier findings in the second-language-learning literature (e.g., Tinkham, 1993, 1997).

Thus, a consistent pattern that emerged across both experiments was that manipulations that increased difficulty during learning (i.e., a larger set size in Experiment 1 and English-to-French translation direction in Experiment 2) were associated with reduced performance during learning. These effects were large; however, the disadvantage for more difficult learning conditions was eliminated during relearning and the final test in each experiment. These findings are partially consistent with the prior research on desirable difficulties, which generally show a reduced performance during learning but superior performance on delayed testing. Together, these findings show that not all manipulations that introduce difficulty appear to produce long-lasting benefits, or long-lasting harm.

There are limitations to the present study. As mentioned previously, Experiment 1 did not control for spacing, and it is possible that larger set sizes could improve retention if spacing was controlled. If so, then block size could be considered a desirable difficulty when learning second-language vocabulary, as long as spacing is controlled. In Experiment 2, a delayed retention test was not administered for the SC and DC items; thus, it is possible that the SC items, which were more difficult to learn during the second session, may have ultimately been better retained. Regardless of the type of difficulty manipulation, future research should investigate how difficulty influences other aspects of learning. For example, increasing difficulty is presumably perceived as requiring increased effort, and increased effort can be perceived as an indication of lesser learning (de Bruin, 2023) and can also result in decreased motivation.

Author Contributions: Conceptualization, A.F.H. and V.I.S.; methodology, A.F.H. and V.I.S.; software, J.A.K.; validation, J.A.K., A.F.H., and V.I.S.; formal analysis, V.I.S., A.F.H. and J.A.K.; investigation, V.I.S. and A.F.H.; resources, A.F.H.; data curation, J.A.K.; writing—original draft preparation, A.F.H.; writing—review and editing, J.A.K., A.F.H. and V.I.S.; visualization, J.A.K.; supervision, V.I.S.; project administration, A.F.H. and V.I.S.; funding acquisition, A.F.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Aeronautics and Space Administration Grant NNX14AB75A to the University of Colorado Boulder.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Human Research Committee of the University of Colorado Boulder (Protocol Number 1098.14) on [1 August 1999].

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data are available upon request from the authors.

Acknowledgments: We are indebted to Michelle Vette, who conducted Experiment 2 as her honors thesis. We are also grateful for the thoughtful comments and guidance provided by two anonymous reviewers of this paper, including many references to the relevant literature.

Conflicts of Interest: The authors declare that there is no conflict of interest.

Appendix A. Items Used in Experiment 1

Category	*French*	*English*
body parts	jambe	leg
	dos	back
	bouche	mouth
	figure	face
	yeux	eyes
	doigt	finger
clothing	manteau	coat
	chemise	shirt
	chaussures	shoes
	cravate	tie
	jupe	skirt
	lunettes	glasses
dining	fourchette	fork
	cuiller	spoon
	couteau	knife
	assiette	plate
	verre	glass
	serviette	napkin
food	pain	bread
	beurre	butter
	lait	milk
	oeufs	eggs
	jambon	ham
	fromage	cheese
school	cahier	notebook
	livre	book
	stylo	pen
	ordinateur	computer
	devoir	assignment
	note	grade
transportation	voiture	car
	velo	bicycle
	avion	airplane
	moto	motorcycle
	camion	truck
	bateau	boat

Appendix B. Items Used in Experiment 2

Category	*French*	*English*
body parts	bouche	mouth
	bras	arm
	coeur	heart

	cou	neck
	doigt	finger
	dos	back
	genou	knee
	jambe	leg
	main	hand
	yeux	eyes
	oreille	ear
	figure	face
clothing	calecon	shorts
	chapeau	hat
	chaussures	shoes
	chemise	shirt
	gant	glove
	lunettes	glasses
	manteau	coat
	robe	dress
	cravate	tie
	veston	jacket
	jupe	skirt
	tailleur	suit
dining	jatte	bowl
	assiette	plate
	couteau	knife
	nappe	tablecloth
	tasse	cup
	fourchette	fork
	cuiller	spoon
	verre	glass
	louche	ladle
	serviette	napkin
	timbale	mug
	cruche	pitcher
food	aubergine	eggplant
	beurre	butter
	epices	spices
	farine	flour
	fromage	cheese
	gateau	cake
	haricots	beans
	legume	vegetable
	nouille	noodle
	oeufs	eggs
	pain	bread
	cerise	cherry
people	epoux	spouse
	fille	girl
	fils	son

	garcon	boy
	homme	man
	soeur	sister
	boulanger	baker
	ecrivain	writer
	femme	women
	pretre	priest
	reine	queen
	roi	king
places	bibliotheque	library
	ecole	school
	eglise	church
	epicerie	grocery
	magasin	store
	maison	house
	piscine	pool
	plage	beach
	gare	station
	ferme	farm
	immeuble	apartment
	salle	room
school	valise	bag
	cahier	notebook
	livre	book
	note	grade
	ordinateur	computer
	stylo	pen
	craie	chalk
	crayon	pencil
	devoir	assignment
	pupitre	desk
	chaise	chair
	grattoir	eraser
transportation	bateau	boat
	camion	truck
	chaloupe	canoe
	fusee	rocket
	velo	bicycle
	velomoteur	moped
	voiture	car
	cheval	horse
	planeur	glider
	avion	plane
	charrette	wagon
	patin	skate

References

Bach, N. T. A., & Barclay, S. (2025). The effect of semantic and physical similarity on vocabulary learning. *The Language Learning Journal*, *53*(1), 40–52. [CrossRef]

Battig, W. F. (1972). Intratask interference as a source of facilitation in transfer and retention. In R. F. Thompson, & J. F. Voss (Eds.), *Topics in learning and performance* (pp. 131–159). Academic Press.

Battig, W. F. (1979). The flexibility of human memory. In L. S. Cermak, & F. I. M. Craik (Eds.), *Levels of processing and human memory* (pp. 23–44). Erlbaum.

Bialystok, E. (2011). Reshaping the mind: The benefits of bilingualism. *Canadian Journal of Experimental Psychology*, *65*(4), 229–235. [CrossRef]

Bialystok, E. (2021). Bilingualism: Pathway to cognitive reserve. *Trends in Cognitive Sciences*, *25*(5), 355–364. [CrossRef] [PubMed]

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition*, *3*(3), 165–170. [CrossRef]

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, *9*(4), 475–479. [CrossRef]

Bjork, R. A., & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *American Journal of Psychology*, *128*, 241–252. [CrossRef]

Bolger, P. A., & Zapata, G. C. (2011). Semantic categories and context in L2 vocabulary learning. *Language Learning*, *61*(2), 614–646. [CrossRef]

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. [CrossRef]

Carpenter, S. K. (2017). Spacing effects on learning and memory. In J. T. Wixted (Ed.), *Learning and memory: A comprehensive reference* (2nd ed., pp. 465–485). Academic Press. [CrossRef]

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*(5), 671–682. [CrossRef]

Carvalho, P. F., & Goldstone, R. L. (2014). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology*, *5*, 936. [CrossRef]

Chan, J. C. K., Davis, S. D., Yurtsever, A., & Myers, S. J. (2024). The magnitude of the testing effect is independent of retrieval practice performance. *Journal of Experimental Psychology: General*, *153*(7), 1816–1836. [CrossRef] [PubMed]

Currie, N. K., & Muijselaar, M. M. L. (2019). Inference making in young children: The concurrent and longitudinal contributions of verbal working memory and vocabulary. *Journal of Educational Psychology*, *111*(8), 1416–1431. [CrossRef]

Czyż, S. H., Wójcik, A. M., Solarská, P., & Kiper, P. (2024). High contextual interference improves retention in motor learning: Systematic review and meta-analysis. *Scientific Reports*, *14*, 15974. [CrossRef]

de Bruin, A. B. H. (2023). Dealing with desirable difficulties: Supporting students to accept, reduce, or silence effort. *Medical Science Educator*, *33*(Suppl. 1), 5–9. [CrossRef] [PubMed]

Eglington, L. G., & Kang, S. H. K. (2017). Interleaved presentation benefits science category learning. *Journal of Applied Research in Memory and Cognition*, *6*(4), 475–485. [CrossRef]

Griffin, G. F., & Harley, T. A. (1996). List learning of second language vocabulary. *Applied Psycholinguistics*, *17*(4), 443–460. [CrossRef]

Hashemi, M. R., & Gowdasiaei, F. (2005). An attribute-treatment interaction study: Lexical-set versus semantically-unrelated vocabulary instruction. *RELC Journal*, *36*(3), 341–361. [CrossRef]

Healy, A. F., Fendrich, D. W., Crutcher, R. J., Wittman, W. T., Gesi, A. T., Ericsson, K. A., & Bourne, L. E., Jr. (1992). The long-term retention of skills. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 87–118). Erlbaum.

Hoshino, Y. (2010). The categorical facilitation effects on L2 vocabulary learning in a classroom setting. *RELC Journal*, *41*(3), 301–312. [CrossRef]

Jongman, S. R., Khoe, Y. H., & Hintz, F. (2021). Vocabulary size influences spontaneous speech in native language users: Validating the use of automatic speech recognition in individual differences research. *Language and Speech*, *64*(1), 35–51. [CrossRef]

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103. [CrossRef]

Kievit, R. A., Hofman, A. D., & Nation, K. (2019). Mutualistic coupling between vocabulary and reasoning in young children: A replication and extension of the study by Kievit et al. (2017). *Psychological Science*, *30*(8), 1245–1252. [CrossRef] [PubMed]

Kievit, R. A., Lindenberger, U., Goodyer, I. M., Jones, P. B., Fonagy, P., Bullmore, E. T., & Dolan, R. J. (2017). Mutualistic coupling between vocabulary and reasoning supports cognitive development during late adolescence and early adulthood. *Psychological Science*, *28*(10), 1419–1431. [CrossRef] [PubMed]

Kole, J. A., & Healy, A. F. (2007). Using prior knowledge to minimize interference when learning large amounts of information. *Memory & Cognition*, *35*(1), 124–137. [CrossRef]

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*(9), 1297–1317. [CrossRef]

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592. [CrossRef]

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review*, *27*(6), 1139–1165. [CrossRef]

McNamara, D. S., & Healy, A. F. (2000). A procedural explanation of the generation effect for simple and difficult multiplication problems and answers. *Journal of Memory and Language*, *43*(4), 652–679. [CrossRef]

Mielicki, M. K., & Wiley, J. (2022). Exploring the necessary conditions for observing interleaved practice benefits in math learning. *Learning and Instruction*, *80*(3), 101–115. [CrossRef]

Mousavi, S. H., Kakhki, A. S., Fazeli, D., Vogel, L., Horst, F., & Schöllhorn, W. I. (2024). Effects of contextual interference and differential learning on performance and mental representations in a golf putting task. *European Journal of Sport Science*, *24*(3), 289–301. [CrossRef]

Nakata, T., Suzuki, Y., & He, X. (2023). Costs and benefits of spacing for second language vocabulary learning: Does relearning override the positive and negative effects of spacing? *Language Learning*, *73*, 799–834. [CrossRef]

Nakata, T., & Webb, S. A. (2016). Does studying vocabulary in smaller sets increase learning? The effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition*, *38*, 523–552. [CrossRef]

Olinghouse, N. G., & Leaird, J. T. (2009). The relationship between measures of vocabulary and narrative writing quality in second-and fourth-grade students. *Reading and Writing: An Interdisciplinary Journal*, *22*(5), 545–565. [CrossRef]

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing: An Interdisciplinary Journal*, *26*(1), 45–65. [CrossRef]

Paap, K. R., Johnson, H. A., & Sawi, O. (2015). Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances. *Cortex*, *69*, 265–278. [CrossRef]

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, *26*(6), 462–483. [CrossRef] [PubMed]

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481–498. [CrossRef]

Sana, F., Yan, V. X., Kim, J. A., Bjork, E. L., & Bjork, R. A. (2018). Does working memory capacity moderate the interleaving benefit? *Journal of Applied Research in Memory and Cognition*, *7*(3), 361–369. [CrossRef]

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (1998). Contextual interference effects in foreign language vocabulary acquisition and retention. In A. F. Healy, & L. E. Bourne Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 77–90). Erlbaum.

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*, 419–440. [CrossRef]

Schneider, V. I., Healy, A. F., Ericsson, K. A., & Bourne, L. E., Jr. (1995). The effects of contextual interference on the acquisition and retention of logical rules. In A. F. Healy, & L. E. Bourne Jr. (Eds.), *Learning and memory of knowledge and skills: Durability and specificity* (pp. 95–131). Sage.

Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, *5*, 772. [CrossRef]

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(2), 179–187. [CrossRef]

Sorensen, L. J., & Woltz, D. J. (2016). Blocking as a friend of induction in verbal category learning. *Memory & Cognition*, *44*(7), 1000–1013. [CrossRef]

Terai, M., Yamashita, J., & Pasich, K. E. (2021). Effects of learning direction in retrieval practice on EFL vocabulary learning. *Studies in Second Language Acquisition*, *43*(5), 1116–1137. [CrossRef]

Tinkham, T. (1993). The effect of semantic clustering on the learning of second language vocabulary. *System*, *21*, 371–380. [CrossRef]

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research*, *13*, 138–163. [CrossRef]

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2010). Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. *Quarterly Journal of Experimental Psychology*, *64*(3), 447–466. [CrossRef] [PubMed]

Van Overschelde, J. P., & Healy, A. F. (2001). Learning of nondomain facts in high-and low-knowledge domains. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1160–1171. [CrossRef]

Wagner, R. K., & Meros, D. (2010). Vocabulary and reading comprehension: Direct, indirect, and reciprocal influences. *Focus on Exceptional Children*, *2010*(1), 1G1-245473138.

Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal*, *40*(3), 360–376. [CrossRef]

Xie, H., Wang, F., Hao, Y., Chen, J., An, J., Wang, Y., & Liu, H. (2017). The more total cognitive load is reduced by cues, the better retention and transfer of multimedia learning: A meta-analysis and two meta-regression analyses. *PLoS ONE*, *12*(8), e0183884. [CrossRef]

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Review

The Use of Retrieval Practice in the Health Professions: A State-of-the-Art Review

Michael J. Serra [1,*], Althea N. Kaminske [2], Cynthia Nebel [3] and Kristen M. Coppola [4]

[1] Department of Medical Education, School of Medicine, Texas Tech University Health Sciences Center, Lubbock, TX 79430, USA
[2] Department of Surgery, Indiana University School of Medicine, Indianapolis, IN 46202, USA; abauerns@iu.edu
[3] Department of Psychiatry and Behavioral Neuroscience, St. Louis University School of Medicine, St. Louis, MO 63104, USA; cynthia.nebel@health.slu.edu
[4] Department of Psychiatry, Rutgers Robert Wood Johnson Medical School, Piscataway, NJ 08854, USA
* Correspondence: michael.serra@ttuhsc.edu; Tel.: +1-806-743-4900

Abstract

Retrieval practice, or the active recall of information from memory, is a highly effective learning strategy that strengthens memory and comprehension. This effect is robust and strongly backed by research in cognitive psychology. The health professions—including medicine, nursing, and dentistry—have widely embraced retrieval practice as a learning and study tool, particularly for course exams and high-stakes licensing exams. This state-of-the-art review examines the historical development, current applications, and future directions for the use of retrieval practice in health professions education. While retrieval-based learning has long been used informally in these fields, its formal recognition as a scientifically supported study method gained momentum in the early 2000s and then saw a surge in both research interest and curricular adoption between 2010 and 2025. This historical review explores the key factors driving this growth, such as its alignment with assessment-driven education and the increasing availability of third-party study resources that rely on retrieval practice as a guiding principle. Despite its proven benefits for learning, however, barriers persist to its adoption by students, including in the health professions. This article discusses strategies for overcoming these challenges and for enhancing retrieval practice integration into health professions curricula.

Keywords: retrieval practice; health professions education; applied cognitive psychology; self-regulated learning; test-enhanced learning

1. Introduction

Retrieval practice (sometimes called 'the testing effect', 'retrieval-based learning' or 'test-enhanced learning') describes the finding that retrieval, or the process of recalling information from the mind, enhances one's learning and memory. Extensive research in cognitive psychology highlights the effectiveness of retrieval practice for enhancing retention and comprehension (Agarwal et al., 2008; Karpicke, 2017). Practice tests and other assessments not only reinforce learning but also provide valuable feedback to guide restudy (Ariel & Karpicke, 2018).

The health professions—including medicine, nursing, and dentistry—have widely adopted retrieval practice as an effective method for learning and studying, particularly in preparation for high-stakes and licensing exams. This review explores how faculty,

students, learning specialists, licensing bodies and educational resource developers have integrated the use of retrieval practice into health education. We have taken a "state-of-the-art" review approach, presenting a historical record which includes the inconsistent endorsement of retrieval practice in these fields prior to 2010, the significant adoption and proliferation of retrieval practice as a topic of research and as a major method of learning from 2010 to the present (2025) and some potential future directions for research and the use of retrieval practice in the health professions going forward.

In the present review, we use the term "learning" to mean any measurable change in behavior following an event indicating a change in knowledge, skills, or attitude (cf. American Psychological Association, 2018). We recognize the breadth of this definition, but the term is defined in many different ways across fields (Barron et al., 2015), so a definition that covers most versions of the concept will necessarily be vague.

1.1. Retrieval Practice

The first experiments investigating the effects of retrieval practice on memory were conducted by Edwina E. Abbott in 1909 (Abbott, 1909). In her master's thesis, she concluded, "recall is always an aid in the learning process" (p. 25). The topic was reported on intermittently during the 20th century (e.g., Gates, 1917; Glover, 1989; McDaniel & Fisher, 1991; Spitzer, 1939; see also Roediger & Karpicke, 2006b) but saw a resurgence in 2006 when Roediger and Karpicke published, "Test-enhanced learning: Taking memory tests improves long-term retention". They concluded that "[t]esting is a powerful means of improving learning, not just assessing it." (p. 249). In the two decades since, interest in the study and application of retrieval practice has only grown (Pan et al., 2024).

The variety of terms used to refer to this effect reflects the nuance in understanding of the phenomena spearheaded by cognitive psychologists in the past two decades. Early reports on the effect in the 2000s used "the testing effect" or "test-enhanced learning", wherein the emphasis was on the finding that people who were tested on material had improved long-term memory and understanding compared to people who simply re-read the same material (e.g., Agarwal et al., 2008; McDaniel et al., 2007; Roediger & Karpicke, 2006a). Research using the terms "retrieval practice" or "retrieval-based learning", on the other hand, emphasized the process of retrieval and the factors that affect it (prior knowledge, feedback, format, etc.; e.g., Ariel & Karpicke, 2018; Karpicke, 2012). For example, if a learner completes a practice test, but performs poorly on it, would they still benefit from retrieval? Research indicates that the learner will benefit from retrieval practice, independent of their overall performance, so long as their performance is off the floor (Chan et al., 2024). Thus, research in cognitive psychology tends to refer to "retrieval practice" and focuses on the underlying cognitive processes, while "the testing effect" is still commonly used outside of the field to describe the educational context.

The nuanced distinctions within the retrieval practice literature can have important implications for the application of these findings to health professions education. For example, medical students make heavy use of flashcard software and multiple-choice question banks to prepare for course and licensing exams, without much regard for the specific cognitive processes engaged by these different forms of retrieval practice. Whereas flashcards promote repeated exposure to discrete facts, medical licensing exams involve complicated, multiple-choice questions that present case-based vignettes and require deep, applied reasoning to choose the best option among many plausible lures (Kolomitro et al., 2020; Little et al., 2019). Retrieval practice will be most beneficial when the type of retrieval practice matches the processing required on the exam (i.e., "transfer-appropriate processing", see Morris et al., 1977), so engaging in retrieval via practice questions might produce more relevant retrieval practice.

1.2. The Growth of Retrieval Practice in Research and Applied Settings

According to Pan et al. (2024), 1215 peer-reviewed articles addressing "the testing effect", "retrieval practice", or "test-enhanced learning" were published from 1999 to 2022 across disciplines and domains. The popularity of retrieval practice within cognitive psychology, and its robust application to educational settings outside of that field, is likely due to three factors. First, it is a reliable and robust effect. In the work by Roediger and Karpicke (2006a), the effect sizes ranged from $d = 0.31$ to $d = 1.26$. Subsequent meta-analyses of the testing effect and retrieval practice found a reliable medium effect size of testing ($g = 0.50$; Rowland, 2014), medium effect sizes when looking at the transfer of test-enhanced learning ($d = 0.40$; Pan & Rickard, 2018) and a medium effect size in applied classroom settings ($g = 0.50$; Yang et al., 2021). It can be found across a variety of materials, retention intervals and test-types, and it requires a small sample size (20–40 per cell) to detect (Agarwal et al., 2021). Despite some initial evidence to the contrary (Van Gog & Sweller, 2015), retrieval practice benefits learning for complex materials (Karpicke & Aue, 2015; Rawson, 2015), including complicated health professions content such as physiology (J. Dobson et al., 2018), chronic wasting disease (Kriechbaum & Bäuml, 2024) and the aerobic production of adenosine triphosphate (J. L. Dobson et al., 2019). Individual differences do not seem to moderate the effect (cf. de Lima & Buratto, 2024), nor does students' prior level of knowledge for a topic (cf. Buchin & Mulligan, 2023; Cogliano et al., 2019). Put simply, it is a powerful effect that is easy to study. Second, when introduced to a broader audience as "the testing effect", this label provided a simple, yet counterintuitive, explanation for the effect. "Taking a memory test not only assesses what one knows, but also enhances later retention, a phenomenon known as the testing effect." (Roediger & Karpicke, 2006a, p. 249). This framing describes memory phenomena in relatable and accessible terms (i.e., "testing" is a familiar concept to many, but "retrieval" or "recall" may be perceived as cognitive jargon), and turns the assumption that tests are used purely for assessment, rather than learning, on its head. This variation from the norm facilitates a third reason why this effect has received a significant amount of attention outside of just cognitive psychology: many educational fields rely heavily on assessment. For example, as early as 2008, Larsen, Butler and Roediger highlighted the benefits of testing and the ease with which medical education, a field that "has focused extensively on assessment issues", could apply the testing effect to the education of its students (Larsen et al., 2008).

While cognitive psychologists continue to debate and dissect the cognitive processes and phenomena involved in retrieval, that type of nuance is unlikely to be found in the fields in which this research is applied. Thus, while we are reporting on the use of retrieval practice and retrieval-based learning in the health professions, we include papers that refer to the testing effect or test-enhanced learning, as well as earlier papers and other publications which used none of those terms but did involve forms of retrieval practice such as practice questions (especially in the health professions).

1.3. Why Have the Health Professions Embraced Retrieval Practice?

Several aspects of education in the health professions have made them a "best-case scenario" for applying retrieval practice and seeing clear benefits from its use. First, the health professions all require students to learn a high volume of very difficult content in a relatively short amount of time, while juggling other personal and professional demands (e.g., trying to gain research experience in a laboratory outside of curricular responsibilities), all while under intense pressure to obtain high scores in courses and to earn high or passing scores on national licensing exams (e.g., Rogers et al., 2016). Second, students of the health professions are typically very highly motivated to perform well in their courses and on licensing exams and are often seeking ways to improve their performance or

make their studying more efficient (e.g., Lawrence et al., 2023; Matsko & Cervantes, 2024). Third, despite the very high admission standards for study in the health professions, a large portion of students begin their training in a health profession lacking strong study skills and habits (e.g., Ahmed et al., 2021; Didarloo & Khalkhali, 2014; Ezeala & Siyanga, 2015; Sisa et al., 2023). Self-regulated learning is a common name for a suite of skills which one uses to be "metacognitively, motivationally, and behaviorally proactive in the learning process" (Zimmerman & Pons, 1986). A common measure of students' skills in self-regulated learning is the Learning and Study Strategies Inventory (LASSI), which includes ten subscales: anxiety, attitude, concentration, information processing, motivation, selecting main ideas, self-testing, test strategies, time management, and using academic resources (Weinstein et al., 1987; Weinstein et al., 2002). One study found half of medical students to be below the 50th percentile on several of the LASSI subscales (O'Sullivan et al., 2024). As well, some studies have found significant correlations between medical students' scores on the LASSI subscales of anxiety, information processing, motivation, selecting main ideas, and test strategies with their performance on biomedical sciences content, anatomical sciences content, and on national licensing exams (Khalil et al., 2017, 2018, 2020). Fortunately, study skills interventions can improve students' study skills as measured by the LASSI (Sisa et al., 2023). Fourth, although retrieval practice is just one tool that students of the health professions can utilize to maximize their learning, it is fairly simple to implement (by students, faculty and software) while being highly effective for enhancing memory and comprehension while reducing forgetting over time (e.g., Agarwal et al., 2021; Ariel & Karpicke, 2018; Donker et al., 2022). It also provides objective feedback on what information the students know or do not know (e.g., Ariel & Karpicke, 2018; England & Serra, 2012; Serra & Ariel, 2014), making retrieval practice a valuable component of self-regulated learning.

Most health profession curriculums include a mix of classroom-based instruction, seminars, laboratory/practical courses and—especially in the later phases of training—hands-on experience treating actual patients (Alishahedani et al., 2019; Barrison et al., 2024). These programs also formally and informally involve a heavy use of "third-party" resources: the written summaries, instructional videos, practice question banks, study software and other learning materials created by publishing companies, test-preparation companies and health education experts to supplement the curriculum. These third-party resources are sometimes referred to as the "parallel curriculum" in the health professions (Barrison et al., 2024). Students of the health professions often become overwhelmed by "resource fatigue" (Alishahedani et al., 2019). Besides the large volume of educational resources available to them, students in these fields must also navigate complicated (and potentially conflicting) advice from faculty, learning specialists, peers, tutors and online "experts" about these resources and ways to study effectively.

Although effectively using retrieval practice is just one component of self-regulated learning, it exemplifies the kind of evidence-based strategy that supports the development of lifelong learning skills—skills essential for success in the health professions. Given the arduous complexity, pace and volume of health education, students must be proactive, reflective and strategic in how they learn. Retrieval practice aligns with these demands: it not only strengthens memory and understanding but also fosters greater metacognitive awareness, helping students identify gaps in their knowledge and adjust their studying accordingly. As such, its integration into health professions education supports not just academic success, but also the formation of adaptive, self-directed professionals prepared for the demands of continuous learning throughout their careers (Artino et al., 2015).

1.4. The Present Review

A state-of-the-art review is a chronological overview of the evolution of knowledge about a phenomenon, highlighting current understanding, the historical developments that shaped it and future research directions. It synthesizes key turning points in the field to offer a modern, forward-looking perspective on the topic, informed by the past and present understanding of the topic (Barry et al., 2022; Grant & Booth, 2009). Accordingly, our review discusses the past, present and future use of retrieval practice as a learning and study method in the health professions.

The health professions have strongly embraced and applied findings from cognitive psychology (including retrieval practice and several other empirically supported methods for improving learning) for the training of their students. These fields are therefore not only an informative case study for how the science of learning can be successfully applied to actual curriculum in an effective way, but can also provide blueprints and best practices for other fields that could benefit from a stronger embrace of empirically supported methods for training students. While the science of learning has been adopted strongly within the health professions, we also note areas for improvement throughout the review, as students (regardless of field of study) are sometimes hesitant to adopt more effective learning strategies compared to less effective ones that they are more familiar with, or they might misunderstand how best to implement the more effective strategies.

We had two major goals for this review. First, to describe and acknowledge the major impact that research on retrieval practice has had on education in the health professions, summarize the efficacy of retrieval practice for learning in this context and highlight how these fields have been using retrieval practice as a learning and study method for the past two decades. Second, to highlight some challenges and future directions for the effective use of retrieval practice in the health professions, many of which are ripe for further empirical study by cognitive psychologists, health professions faculty and even students of the health-professions. We followed Barry et al.'s (2022) six-stage methodology for conducting a state-of-the-art review.

Analytical Approach

In Stage 1 (Research Question), we determined our research question and field of inquiry: How have the health professions used retrieval practice to benefit their students? This question and our approach to answering it were guided by our mutual backgrounds as cognitive psychologists and medical education learning specialists (i.e., "MELS").

In Stage 2 (Timeframe), we determined the period of time that the review would focus on after conducting an overview of the relevant literature. We ultimately chose to organize the review into three important epochs: (1) The (limited) knowledge and use of retrieval practice for learning in the health professions prior to 2010, (2) the embrace and proliferation of retrieval practice by health professions educators from 2010 to 2025 and (3) future directions for retrieval practice in the health professions after 2025.

In Stage 3 (Finalization), we refined the research question and these time frames from our initial assumptions as follows. We initially expected knowledge of retrieval practice in the health professions to be low prior to 2006, when it was highlighted in the cognitive psychology literature by Roediger and Karpicke (2006a). Instead, as our literature review progressed, we recognized that retrieval practice was used heavily in these fields prior to this point, even if it was not labeled as such or its benefits for learning were not fully understood. This status changed greatly by 2010—when it became better understood and identified by name in these fields—so we set our first demarcation to 2010. There was little need for further delineation between 2010 and 2025, as research, understanding and

endorsement of retrieval practice in these fields has only grown in that time. Most relevant empirical studies in these domains appeared in the later portion of this period.

In Stage 4 (Search Strategy), we identified articles to include in our review. We first conducted searches for articles containing the phrases "retrieval practice", "test-enhanced learning" or "testing effect", in combination with phrases such as "health professions", "medical school", "nursing", etc. We prioritized meta-analyses and systematic reviews, but we also found many individual research papers for some health disciplines which do not yet have relevant review papers. We also needed to identify papers that predated the 2010 demarcation and involved retrieval practice but likely did not include phrases such as "retrieval practice", "test-enhanced learning" or "the testing effect". This included searches for articles involving practice-test questions, problem-based learning and using "pimp" or "pump" questions, all of which are forms of retrieval practice that have been stressed in the health professions but are almost never referred to as "retrieval practice" (especially prior to 2010). We also searched for the occurrence of terms related to retrieval practice in websites, handouts and "third party" resources directed at students or faculty of the health professions to determine how widely this topic has been embraced and endorsed in these fields over the years and at present.

We began Stage 5 (Analyses) by reading articles we initially identified in our search, familiarizing ourselves with the literature to date and noting similarities and differences across articles. We also noted gaps in knowledge or understanding, especially if the gaps suggested viable research questions. As noted, this stage included a correction of our initial assumption that retrieval practice was not widely used in the health professions prior to 2006, albeit without such labeling and without knowledge of the benefits of retrieval for learning per se; most early sources viewed forms of retrieval practice such as answering practice test questions for practice and assessment purposes only.

The authors of this review are all cognitive psychologists with unique backgrounds and experiences in research and teaching across psychology and medical education. All currently serve as medical education learning specialists at their respective medical schools. Their shared experience working closely with medical students in individualized instructional settings may have shaped their perspectives when interpreting literature on retrieval practice. For Stage 6 (Reflexivity), we offer the following reflection: although the present review reflects our expertise as cognitive psychologists and as medical–education learning specialists, our interpretation of the state-of-the-art of retrieval practice in the health professions—as well as suggestions for potential new research questions in these domains—is subjective and likely biased by our training and current professional roles. For example, we do not bring to this situation the opinions, experience or expertise of professionals or students of the health professions.

2. The Inconsistent Use of Retrieval Practice in the Health Professions (Before 2010)

For the purposes of this review, we will discuss the health professions (i.e., medicine, nursing, dentistry, pharmaceuticals, etc.) as a conglomerate, only noting specific articles, findings or trends from a specific domain when necessary. Though these disciplines all differ from each other in many important ways, in the context of this review they have much in common. As we have already noted, they all require students to learn a high volume of very difficult content in a relatively short amount of time, and they all involve high-stakes exams during the curriculum culminating in high-stakes national licensing exams. There is a clear need for students in all these disciplines to use effective and efficient studying strategies to progress through their curriculum in a timely manner. Obviously, students in these disciplines were studying effectively under these pressures well before the formal

recognition of the benefits of retrieval practice for learning by cognitive psychologists in the early 2000s, but how were they doing it, and to what extent did their methods involve retrieval practice?

2.1. The Fine Art of "Pimping" Students

There is at least one classical technique in the health professions that clearly emulates retrieval practice as a learning and diagnostic tool, and that is "pimping" (Antonoff & D'Cunha, 2011). This term derives from the German word "Pümpfrage", which is often translated to English as "pimp questions", although "pump questions" is a more correct translation (Brancati, 1989). The technique was first described by Harvey in 1628 England and later popularized by Koch in 1889 Germany. This method involves an attending physician (or other person of greater knowledge, such as a senior medical student) rapidly questioning a resident or student of lesser knowledge on complex topics. It is frequently employed during hospital rounds, surgical procedures and academic discussions, and typically occurs in front of peers and patients (Anderson, 2013; Detsky, 2009). By design, the technique is meant to create a high-pressure retrieval situation somewhat like a medical emergency (Brancati, 1989). But it is also meant to encourage learning through negative emotions like fear and intimidation (Gooding et al., 2017; Musselman et al., 2005) in an experience that is embarrassing and borders on traumatic (Jimenez et al., 2024). Some have called for a change in label to this method to avoid an association to pimping/prostitution while still maintaining a negative connotation, including "improper Socratic method" or "malignant questioning" (cf. Nagarur et al., 2019).

Pimping students is often referred to as an "art" by faculty and students in the health professions (Brancati, 1989; Detsky, 2009). Several survey-based studies indicate that faculty and students perceive pimping to be an effective technique for learning and for identifying unknown information (e.g., Abou-Hanna et al., 2021; McEvoy et al., 2019; Rucker et al., 2023; Wear et al., 2005). Given what we currently know about the benefits of retrieval practice for learning, questioning students in this way should enhance their learning and understanding of the information queried and help them identify gaps in their knowledge (cf. Antonoff & D'Cunha, 2011). To our knowledge, however, no study has demonstrated objective improvements in learning from this method (cf. Jimenez et al., 2024; McCarthy & McEvoy, 2015; Reifler, 2015).

Although pimping is still practiced in the health professions today, many articles have questioned the lack of efficacy data and suggested that the negative effects of this practice on students might outweigh any positive effects it might have for their learning. Many faculty and students perceive the practice to be abusive, traumatic and a form of bullying (cf. McCarthy & McEvoy, 2015; Reifler, 2015; Wear et al., 2005). Unironically, Jimenez et al. (2024) hypothesized that the process of pimping could aid students' learning via the same mechanisms that lead to invasive "flashback" memories in post-traumatic stress disorder. In addition, berating residents and medical students in front of patients and their family members for lacking knowledge might create distrust between health professionals and patients and can stifle communication and deliberation about patient care within teams of health professionals, putting patients at risk (cf. Anderson, 2013).

The process of "pimping" students in the health professions by asking them rapid-fire, difficult questions under social pressure is a pedagogically and ethically questionable practice, but it likely represents an example of using retrieval practice to teach and learn information in the health professions prior to 2010. As with some other forms of retrieval practice utilized in these fields, it was not usually referred to as such before 2010 (in fact, "pimping" is still not often framed as a form of retrieval practice even in 2025; but see Antonoff & D'Cunha, 2011). In contrast to some of the other forms of retrieval practice we

discuss in this review, however, the presumed benefits of pimp questions for memory and understanding per se were more often better acknowledged even prior to 2010. That said, we could find no empirical studies that have compared the effects of "pimping" to any other learning modality in the health professions, even lecturing.

Some experts contrast pimping with a classic form of discovery learning known as Socratic questioning. Although both techniques involve teaching by questioning rather than by information delivery, Socratic questioning involves challenging students' current beliefs to help them recognize their misconceptions and correct their knowledge through critical reflection (Stoddard & O'Dell, 2016; Yudcovitch & Hayes, 2014). Whereas pimping embarrasses or degrades the learner, Socratic questioning requires psychological safety, allowing learners to focus on increasing their knowledge and skill rather than fending off threats to their self-worth (Stoddard & O'Dell, 2016). This method, however, might not produce demonstrable benefits over simple lecturing (e.g., Yudcovitch & Hayes, 2014).

2.2. Other Unrealized Forms of Retrieval Practice

Prior to the 2010s, students of the health professions—much like those today—used a combination of information intake (i.e., learning from lectures, textbooks, laboratory activities and treating simulated and actual patients) and forms of retrieval practice (e.g., flashcards, practice test questions, writing summaries, drawing diagrams, working with a study partner, etc.) to learn and practice the content in their curriculum (Thadani et al., 2000). Based on education research and expert guidance at the time, however, it seems that the emphasis was more squarely on information intake as the primary method of learning, but forms of retrieval such as answering practice-test questions were intended only as practice and assessment, not as a learning strategy themselves. For illustrative purposes, let us consider guidance and related research prior to 2010 for how medical students can prepare for the United States Medical Licensing Examination (USMLE) Step 1, the first level of licensing exam that medical students in the United States must pass (typically taken in the second year of medical school).

Problem-based learning (PBL) was a common alternative to the "standard" medical curriculum during this time (Albanese & Mitchell, 1993). Whereas the common approach favored presenting information first and having students apply it later, PBL exposed students to realistic problems and situations before they had learned all the information necessary to do so (e.g., a hypothetical patient presents with given symptoms; students must decide the diagnosis and potential treatment). The intent was to nudge students into an information-seeking mode, where they are responsible for determining what they need to learn for a problem they cannot yet solve (Barrows, 1986). By 2000, medical education articles examining the efficacy of a PBL approach for passing Step 1 focused on its use as a remediation tool for struggling students and found that PBL improved outcomes for those students (e.g., Walters et al., 1999; see also Blake et al., 2000). As with other guidance prior to 2010, PBL articles often discuss related benefits for learning, such as identifying incomplete or incorrect understanding of topics, and building general skills at answering Step 1 questions (e.g., Walters et al., 1999). Although the PBL approach naturally involves students doing many practice questions ahead of Step 1, guidance for this approach did not include the direct benefits of retrieval practice per se on learning.

Similarly, "The Step 1 Method" (Gebremedhin, 2012), a popular study guide for this exam, relies heavily on the use of practice questions as a tool for preparing for the exam. In this case, the primary focus is on using questions for information exposure, repetition, identifying unknown content and practicing and improving test-taking speed and broad question-answering skills. At least in its handbook form (Gebremedhin, 2012), however, this method does not acknowledge the benefits of retrieval practice for learning. Other

similar Step 1 courses demonstrated utility for addressing knowledge gaps for students who struggled in the medical curriculum (Thadani et al., 2000), but typically focused on content presentation, review and broad study/test strategies, without any mention of practice questions in a retrieval-practice way (Zhang et al., 2005).

In a research article aimed at MELS and other academic counselors, Coumarbatch et al. (2010) investigated several factors that might predict a student's likelihood of passing or failing the Step 1 exam on their first attempt. They also suggested several potential interventions for remediation, organized into three levels of severity. Although the specifics of these interventions are sparse, only the intervention designed for the students most at-risk of failing specifically mentions completing practice questions. This aspect of the intervention is called the "3500 Question Challenge". Presumably, amongst the other things listed (i.e., mandatory academic skills counseling, tutoring, back-to-school reviews, case studies, etc.), these students would have also been expected to complete 3500 practice Step 1 questions. This number might seem daunting to those outside of medical education or the other health professions, but as of this writing (2025), this is the approximate number of questions in each of the two most popular Step 1 question banks (e.g., UWorld and AMBOSS). While completing thousands of practice questions can be beneficial as it promotes retrieval practice, we note that guidelines that focus solely on the number of practice questions to be completed lack nuanced recommendations for how students should be approaching the questions, receiving feedback and reviewing the questions to make the most out of these retrieval practice opportunities.

In contrast, in an early guidebook of study techniques for medical students, Saks et al. (1998) alluded to retrieval practice in a variety of forms within the framework of acquisition, maintenance and proficiency. The authors stressed the importance of cumulative review of concepts using small units of information at regular intervals. For example, one retrieval technique suggested to maintain learning for anatomy was sketching out pathways for cranial nerves from memory, which requires retrieval. Similarly, in learning biochemical pathways, the authors suggest that students draw the pathways from memory to test their knowledge. In the proficiency phase, use of practice questions was encouraged at regular intervals between examinations and well before the student felt ready for the examination. The main objective in their promotion of practice questions was to identify weak topics and ensure that application of knowledge would be successful from initial acquisition to testing. The authors state, "remember that you do not have to reach proficiency to begin questions, as this is a process that can be helpful in deciding what your focus should be" (p. 252). Similarly, in preparation for licensing exams, questions for self-testing were described as an effective way to review information, assess knowledge and prioritize follow-up study. The authors do not use the terms "retrieval practice" or "testing effect" but consistently support that recall is highest when regular review of information that is organized in a meaningful way is completed and use of practice questions early and often are stressed throughout their text. Importantly, we must acknowledge that this guidebook identified and anticipated many of the benefits of retrieval practice for studying and learning—based on the authors' experience advising medical students for academic success—that are now taken as fact after decades' worth of research in cognitive and educational psychology.

In summary, the dominant educational emphasis in the health professions prior to the 2010s was on content exposure rather than retrieval as a core learning mechanism. Health professions students typically focused on a combination of information intake—via lectures, textbooks and clinical experiences—but practice questions were reserved for assessment of progress. Problem-based learning (PBL) was a notable alternative approach that encouraged active information seeking but also lacked explicit recognition of the learning benefits of retrieval practice. Popular study guides and courses that prepared

students for licensing exams similarly emphasized question volume without describing the benefits of engaging with the questions (a tendency that persists in many fields besides the health professions; cf. Hartwig & Dunlosky, 2012). An exception was the early work of Saks et al. (1998), who, without using terms like "retrieval practice," described study methods that strongly aligned with its principles—such as recalling information from memory, cumulative review and early use of practice questions to guide study.

2.3. Research on "The Testing Effect" Begins

Besides guidance in books, curriculum and study guides, there are also examples of health professions research examining the effects of retrieval practice on learning prior to its formal recognition and labeling in cognitive psychology (i.e., Roediger & Karpicke, 2006a). For example, Van Hoof et al.'s (2023) review of retrieval practice studies in nursing includes several studies that predate this time point: two from 2001 and one from 2006. Trumble et al.'s (2024) review includes some studies by Logan et al. (1975), who examined massed practice (studying a topic in one large "block" of time) versus distributed or spaced practice (studying a topic repeatedly or in sections spaced over time, often with the study of other content interspersed) in the acquisition of physical dentistry skills. Although distributed or spaced study does not explicitly involve retrieval practice, it likely benefits from some of the same mechanisms as both require the reactivation of information at a delay.

Shortly after the modern, formal recognition of the benefits of retrieval practice for learning in cognitive psychology (i.e., by Roediger & Karpicke, 2006a), several articles appeared by Larsen, Butler, Roediger and others (e.g., Larsen et al., 2008, 2009, 2013, 2015) indicating a clear partnership of cognitive psychology and medical education researchers examining retrieval practice in a medical education context. In the first such article, Larsen et al. (2008), suggested that, "If test-enhanced learning were to be implemented in medical education, the focus of testing and assessment might shift significantly. Tests would no longer be considered neutral tools of measurement, but rather active instruments to aid in the acquisition and retention of knowledge." (p. 963). Although brief, their paper and its framing clearly suggest that retrieval practice per se was not yet explicitly endorsed in medical education (see also Larsen et al., 2009). Presumably, it was not yet recognized in the other health professions either. To Larsen et al.'s credit, when they suggested that tests would become "active instruments to aid in the acquisition and retention of [medical] knowledge.", they accurately predicted what would soon become the norm in the health sciences less than fifteen years later: students using various forms of retrieval practice not just as a form of assessment but also as a form of learning.

3. Recognition, Verification, and Adoption of Retrieval Practice in the Health Professions (2010–2025)

By the 2010s, knowledge of the benefits of retrieval practice had clearly reached the health professions. Initiated by studies and papers involving both cognitive psychology researchers and health professions researchers (e.g., Larsen et al., 2008, 2009), the historical record indicates a rapid proliferation of papers examining the effects of retrieval practice in authentic, health profession contexts (especially in the last decade). For example, Deng et al. (2015) found that medical students frequently use self-initiated retrieval practice (e.g., use of digital flashcard products), with or without spaced or distributed repetition, to prepare for the Step 1 licensing exam, and that such retrieval practice enhances learning and test performance. Metz et al. (2021) found that spaced or distributed retrieval practice improved year one dental students' learning outcomes. Studies by Giuliodori et al. (2008) and Kleinsorgen et al. (2018) indicate that retrieval practice aids the learning of veterinary medicine content. Pre-lecture quizzes can enhance students' memory for neuroanatomy

and physiology terminology and concepts (Pan et al., 2019). Articles also support the use of retrieval practice while studying for the Medical College Admission Test (MCAT), a standardized examination for prospective medical students commonly used in the United States and in some other countries (e.g., Steed & Kadavakollu, 2019; Swan Sein et al., 2020).

Several review papers summarize findings from studies within a specific health domain (e.g., just nursing), across multiple health domains, or along with studies from many other diverse fields (see Table 1 for some examples). Importantly, most studies and meta-analyses demonstrate positive benefits for retrieval practice in health professions education. Green et al.'s (2018) review across the health professions found robust benefits of retrieval practice for learning, including effects that extend beyond course examinations to clinical applications. Trumble et al. (2024) also found that retrieval practice was effective at improving academic performance in the health professions in the majority of the studies they reviewed. Thompson and Hughes (2023) argued that retrieval practice helps the learning of radiology content, but their review only identified eight articles that met criteria, with just five finding a positive benefit of retrieval practice. Goldman et al. (2024) found that the use of Anki flashcard software (presumably, Version 2.0 or 2.1 given the timeline) was positively associated with exam outcomes in the health professions. Wu et al. (2021) identified several factors positively associated with medical students' scores on the Step 1 exam (before that exam changed to pass/fail in 2022), including some that reflect retrieval practice such as greater use of flashcards and practice questions during study.

Table 1. Examples of recent reviews on the efficacy of retrieval practice for learning that include the health professions.

Discipline	Review Type	Authors/Year	Studies	Study Years
Medicine Only	Systematic	(Wu et al., 2021)	29	2005–2019
Nursing Only	Scoping	(Van Hoof et al., 2023)	25	2001–2021
Radiology Only	Systematic	(Thompson & Hughes, 2023)	8	2014–2022
Health Professions	Systematic	(Green et al., 2018)	19	2009–2016
	Scoping	(Barrison et al., 2024)	64	2011–2024
	Systematic	(Goldman et al., 2024)	8	2015–2022
	Systematic	(Trumble et al., 2024)	63	1975–2022
Health and Others	Meta-Analysis	(Adesope et al., 2017)	3 *	2009–2010 *
	Meta-Analysis	(Pan & Rickard, 2018)	7 *	2009–2013 *
	Systematic	(Moreira et al., 2019)	4 *	2007–2015 *
	Systematic	(Agarwal et al., 2021)	10 *	2009–2017 *
	Systematic	(Rivers, 2021)	6 *	2015–2019 *
	Meta-Analysis	(Yang et al., 2021)	93 *	1968–2019 *

* Note that, for the six reviews that included health profession studies along with non-health profession studies, we have only included information (i.e., number of studies and years of those studies) for those studies that were directly relevant to the health professions.

Most empirical articles examining retrieval practice in a health professions context were published between the years 2010 and 2025. Furthermore, as indicated by several of the reviews in Table 1, most of these articles were published at the more recent end of that range. This pattern suggests that the knowledge and empirical examination of retrieval practice in the health professions only continues to grow. That said, there are some articles in the reviews in Table 1 that appeared before 2010, and even before the publication of the work by Roediger and Karpicke (2006a). Per our discussion of the "inconsistent" use of retrieval practice in the health professions prior to 2010, these data points not only support our earlier suggestion that students of the health professions were using various forms of retrieval practice to aid their learning prior to 2010 (and 2006), but also that researchers in

these domains were studying it in various forms, albeit without the more common labels for the effect that would follow in cognitive psychology such as "the testing effect".

Health professions research involving the effects of retrieval practice also appears in domain-general reviews of the effect (e.g., Moreira et al., 2019). For example, Agarwal et al. (2021) included several papers examining the topic in medicine in their review. Those authors noted that retrieval-practice research in the medical context is often very applied and authentic and that it typically uses rephrased questions rather than repetitions of old questions as is common in some laboratory studies and in some other fields. Of importance here, some cognitive psychology studies have failed to find "far transfer effects" of retrieval practice such that the benefits of retrieval for learning are strongest when the retrieval practice tests the same information as the final criterion test but is weaker or absent when the information tested at both times is less closely related (cf. Wooldridge et al., 2014; Yeo & Fazio, 2019). As well, many such studies in the medical context involved designs where retrieval practice was compared to active, "non-retrieval" comparisons, not simply compared to students not using any form of restudying at all. These are all good markers of quality applied work but also help to ensure that we can trust that retrieval practice is helpful for learning in the health professions context.

3.1. Adoption in Formal Training or Curriculum

It is difficult to fully assess the extent to which retrieval practice is formally endorsed as an effective learning and study method in the health professions. For example, a faculty member or learning specialist might advise a struggling student to use retrieval practice and to complete practice questions while studying, but there would be no public record of that endorsement. Nevertheless, some evidence exists for the formal recommendation of retrieval practice in the health professions across published empirical papers indicating its efficacy, in opinion pieces written by faculty or students, on the websites of schools of the health professions, in supporting documents from third-party learning resources and in online content written by past and present students of the health professions. Table 2 includes several such examples but is certainly not an exhaustive list. Although most of these resources are aimed at the individual student/learner who is likely engaging in self-regulated study on their own, some of these resources are aimed at faculty, who can affect the study behaviors of multiple students at once through course and assignment design or via explicit instruction to students about the utility of retrieval practice for learning.

Larsen et al. (2008) proposed that retrieval practice aligns with the use of assessment to improve medical education and provides a valuable tool to help students to learn and to retain essential knowledge. Thus, some health professions faculty have incorporated retrieval practice through frequent, low-stakes testing (cf. Swan Sein et al., 2021b). Deng et al. (2015) also encouraged medical schools to explore these methods and to research optimal implementation strategies. As the content in Tables 1 and 2 indicate, numerous empirical and editorial articles support including retrieval practice in health professions education. In addition, some schools explicitly discuss and recommend retrieval practice on their websites and in other resources. These resources, which have greatly expanded since 2010, reference key cognitive psychology concepts such as retrieval practice, spaced retrieval and metacognition, citing studies from cognitive and educational psychology as well as health professions research to back those assertions.

Table 2. Example journal, university, and third-party resources explicitly endorsing the use of retrieval practice for learning in the health professions.

Type	Discipline	Source	Title
Journal Article	Anesthesiology	Weidman and Baker (2015), *Anesthesia & Analgesia*	"The Cognitive Science of Learning: Concepts and Strategies for the Educator and Learner"
	Medicine	Cutting and Saks (2012), *Medical Teacher*	"Twelve tips for utilizing principles of learning to support medical education"
	Medicine	Swan Sein et al. (2021a), *Medical Teacher*	"Twelve tips on guiding preparation for both high-stakes exams and long-term learning"
	Medicine	Ahmed et al. (2021), *Medical Teacher*	"Optimizing preclinical learning with retrieval practice: A call to action"
	Medicine	Coppola et al. (2023), *Academic Medicine*	"Determining Root Causes of Poor Academic Performance to Provide Wraparound Support for Preclerkship Medical Students"
University Resource	Medicine	(Loma Linda School of Medicine, 2021)	"The Science (and Illusions) of Learning: Instructional Design Tips for Teachers to Enhance Learners' Long-term Memory Retention"
	Medicine	(The Robert Larner College of Medicine at University of Vermont, n.d.)	"The Science of Learning. Learning Theory: Retrieval Practice"
	Nursing	(UConn School of Nursing—University of Connecticut, n.d.)	"Learning Science Strategies Series, Module 3: Retrieval Practice"
Third-Party Resource	Multiple Fields	(Anki, n.d.)	Anki Manual: "Background"
	Medicine	(AMBOSS, n.d.)	"Internal Medicine (ABIM) Board Review"
	Nursing	(UWorld, 2021)	"Practice Questions: More Important than Ever for NCLEX Success"
	Medicine	(Lecturio, 2024)	"Active Recall and Retrieval Practice in Medical Education"
	Medicine	(Student Doctor Network, 2018)	"Gain a Deeper Understanding with the Power of Test-Enhanced Learning"
	Multiple Fields	(TrueLearn, n.d.)	"Goodbye, Q-Banks. Hello, SmartBanks."

3.2. Adoption in the "Parallel Curriculum"

Table 2 also indicates that we can find the explicit explanation and endorsement of retrieval practice in supporting documents created for the third-party learning resources commonly used in the health professions (e.g., Anki, n.d.; AMBOSS, n.d.; UWorld, 2021), and in blogs and social media posts written by students and others in the health professions. As noted earlier, such writing about how to use those resources is part of the "parallel curriculum" in the health professions (Barrison et al., 2024). Many modern educational resources such as flashcard software have adopted retrieval practice and other findings from cognitive psychology into their design. Students of the health professions also obtain "unofficial" information from their peers, both in-person and online. In their summary of how heavily medical students (and presumably students of the other health professions as well) rely on online forums such as Reddit for study advice, Ronner and Linkowski (2020) suggested that, "On the basis of forum recommendations and advice from their peers, students couple this type of flashcard schedule with a number of commercial Step-prep resources . . . to build a rigorous, months-long study schedule devoid of any input from a brick-and-mortar medical school. . . . [T]his is increasingly the new normal in preclinical medical education." (p. 1330). As such, it is paramount that students receive tips and information embedded into these resources that clearly and accurately describe the benefits and rationale for effective study techniques such as retrieval practice or spaced repetition,

as these resources—and advice from their peers—might be the primary source(s) from which they learn about such techniques.

4. Broadening and Strengthening the Use of Retrieval Practice in the Health Professions (2025 and Beyond)

Retrieval practice is clearly effective at increasing learning and exam performance in the health professions (cf. the reviews in Table 1), and students in these fields are using various forms of retrieval practice as they study. But that does not mean that the issue is settled; there are still barriers to students of the health professions using retrieval practice and using it effectively and efficiently. Accordingly, we have organized the final section of this review into three subsections, reviewing some common barriers to the use of retrieval practice by students of the health professions, interventions that could increase its use in these fields and factors that can make retrieval practice more efficient or more effective. Any of the following topics would benefit from further empirical research by cognitive psychologists and health education researchers.

4.1. Barriers to Retrieval Practice

Despite the numerous benefits of retrieval practice for learning, some students are hesitant to engage in forms of retrieval during the learning process. Other factors simply make it difficult or less appealing for students to do so. In this section, we review some major reasons why students in the health professions might avoid retrieval practice.

4.1.1. Lack of Knowledge About Retrieval Practice

Despite the surge in research related to retrieval practice in the past 15 years, this information does not seem to have reached students and educators more broadly (but see Table 2). Indeed, when asked about the frequency with which they use various strategies, undergraduate students report re-reading at a much higher rate than retrieval practice (Blasiman et al., 2017). Those students who engage in retrieval practice often do so for assessment purposes and not as a learning strategy (Hartwig & Dunlosky, 2012). This may be partially due to educators' inadequate knowledge of retrieval practice as an effective learning strategy as well. Halamish (2018) asked educators to choose the better learning strategy amongst two scenarios, and retrieval practice was only endorsed 30% of the time. More recently, however, Witherby et al. (2025) found that retrieval practice was rated as the most effective strategy by undergraduate psychology students, as well as by educators and parents in the general population, so knowledge of the effect might be increasing. We should note, however, that such studies typically describe behaviors in everyday terms, which respondents either endorse as using (or not), or rate the methods in terms of their presumed efficacy for learning. Such studies rarely, if ever, use cognitive psychology terms such as "retrieval practice" when describing these behaviors for rating, which could dissuade broader spread of the topic. In contrast, the resources listed in Table 2 and others like them typically do use cognitive psychology terms and explanations for their efficacy, perhaps under the assumption that students of the health professions can understand such ideas better or will be more likely to engage in self-testing when framed in a cognitive psychology context than would elementary school children or their parents.

4.1.2. Choosing Less-Effective Learning Strategies

By and large, data from the health professions indicates that students in these areas mimic those in traditional higher education settings: even when students are aware of the benefits of retrieval practice, they tend to choose other less-effective strategies for self-regulated learning (Corazza et al., 2023; Klein et al., 2023; Sheehy et al., 2024). There is far less data to indicate exactly how students engage in retrieval practice during self-regulated

learning, and this is an area ripe for investigation. In one recent study, Sheehy et al. (2024) found that students do not always use retrieval practice in the most effective way. For example, students might choose to look up the answer to a question if they are unsure or have limited knowledge in an area or choose to only engage in additional retrieval for topics that they missed. Both strategies indicate a limited awareness of how retrieval practice is most effective, and result in limited benefits for retrieval practice.

As well, some students seem to be aware but untrusting of the benefits of retrieval practice for learning. While the research is somewhat limited, students in both medical (Furness et al., 2024) and dental education (McAndrew et al., 2015) report choosing re-reading and other less effective strategies more often than retrieval practice. Some of this hesitation might be due to the sources of information that students use when determining how to study. For example, in discussing medical school forums on Reddit, Ronner and Linkowski (2020) noted that students on these forums, "regularly receive and reply to direct messages from users on Reddit asking for the type of advice that a learning professional or study counselor hired by an institution should be providing. The fact that struggling students may be more willing to seek academic help and mentorship from an anonymous online peer than from an educator at their own institution says something meaningful about the alienation students feel from those who are tasked with teaching them." (p. 1330). Ironically, students might interpret their peers' infrequent endorsement of retrieval practice (compared to faculty and learning specialists) as an indicator that it is not an effective strategy.

That said, Kann et al. (2024) recently analyzed the content of the 500 top-ranked r/Step 1 subreddit posts between 26 January 2022 and 10 April 2023. Of the materials most frequently mentioned, the top two were "NBME [National Board of Medical Examiners] materials" (51.40%) and UWorld (40.40%). UWorld is a question bank, and "NBME materials" likely included self-assessment and practice exams that students can take while studying for the Step 1 exam. This may be an indication that students are using retrieval strategies without fully understanding why they are beneficial, which has the potential to lead to "lethal mutations" in the efficacy of the method (Bagley, 2020) due to misperceptions about how retrieval practice "should feel" or the immediate outcomes associated with it. For example, people tend to assume that subjective ease or fluency is an indicator of successful performance, such as effective learning, even when it is not (e.g., Kelley & Lindsay, 1993; Koriat, 2008). In some cases, effortful processing (including effortful retrieval) is associated with better learning or future performance (Benjamin et al., 1998; Serra & Dunlosky, 2005). Students likely need to change how they incorporate experienced fluency into self-evaluations of learning (Briñol et al., 2006).

At the same time, not all recommendations and practices in the health professions are as clearly based on current and accurate empirical knowledge regarding effective studying or learning. While faculty are experts in their respective health professions, they may not be well-versed in cognitive psychology principles that optimize student learning. Consequently, some guidance for faculty continues to promote ineffective methods such as the consideration of learning styles in teaching (e.g., Friedlander et al., 2011) while overlooking empirically supported strategies like retrieval practice and spaced repetition (Augustin, 2014; Selva-Rodriguez & Sandars, 2023; Swan Sein et al., 2021a). Recognizing this gap, the Liaison Committee on Medical Education (LCME) requires structured academic advising that integrates faculty, course directors and other student-support services (see LCME Accreditation Standard 11.1). In response, medical schools have increasingly hired learning experts from fields like cognitive psychology, education and neuroscience to develop and implement evidence-based learning strategies (Rashid et al., 2024). These MELS serve as a bridge between research on learning science and practical application

in medical training, ensuring that faculty and students benefit from proven instructional techniques without requiring all educators to master these principles. By leveraging the expertise of MELS, medical schools can provide faculty (and students) with accurate, research-backed resources while allowing them to focus on their primary role: teaching and mentoring future healthcare professionals.

4.1.3. Misleading Experiences of Retrieval Practice

Students might remain or become hesitant to use retrieval practice while studying because they have previously relied on some forms of retrieval practice that were not appropriate for their learning goals, leading to unimpressive results on assessments. Such experiences can lead students to assume that all forms of retrieval practice are inefficient. For example, students of the health professions make heavy use of flashcard software such as Anki to study for high-stakes exams (e.g., Barrison et al., 2024; Deng et al., 2015). Although flashcard software produces impressive memorization, the quality of learning and future test performance is very dependent on the quality of the flashcards studied and how closely that learning compares to the actual exam(s). Put differently, a student who is effectively memorizing low-level facts via flashcard software might nevertheless struggle on a course or licensing exam where answering questions correctly requires a much more advanced level of comprehension. That student might then assume that all forms of retrieval practice are inefficient for learning, whereas the problem was instead with the type of retrieval practice they utilized. Rather, for students to most effectively transfer their knowledge from retrieval practice to course or licensing exams, it is best if the practice matches the type of processing required on the exam (Barnett & Ceci, 2002; Morris et al., 1977). Students might also wait to use retrieval as a practice assessment shortly before their summative exam to determine their performance. Because students are waiting too long to start using practice questions, they are not maximizing the learning utility of retrieval practice and may not recognize its benefits (cf. Saks et al., 1998). A student in these situations might also assume that all forms of retrieval practice are inefficient and return to less-effective study methods which do not produce a discordant experience (cf. McAndrew et al., 2015; see also Karpicke et al., 2009).

Students often judge their own learning based on the speed and ease (fluency) with which they can retrieve information from memory. In knowledge-based tasks, faster responses are more likely to be correct (Hertwig et al., 2008; Robinson et al., 1997) and considering response speed improves the accuracy of students' judgments of their own knowledge (Koriat & Ackerman, 2010; Serra & Dunlosky, 2005). The experience of fluency, however, can be misleading: the repeated exposure to potential answers also increases response speed and confidence, even if the answers retrieved quickly are incorrect (Kelley & Lindsay, 1993; Koriat, 2008). In a broader sense, attempting retrieval in various forms is often experientially more effortful than many other forms of studying or practice such as re-reading text or re-watching videos that are less likely to produce long-term changes in students' learning. For this reason, students might favor low-return study and restudy techniques over forms of retrieval practice because the former feel easier to engage in.

4.1.4. Negative Beliefs About Testing and Errors

Some students hesitate to use practice questions or other forms of retrieval practice for studying, practice, or self-assessment because they hold negative beliefs about making errors. (cf. Chouvalova et al., 2024; see also Carpenter et al., 2020a, 2020b; Dresel et al., 2013; Pan et al., 2020; Tulis et al., 2017). A well-known topic in psychology and in the health professions is mindset: while a growth mindset refers to the belief that personal traits such as intellectual abilities can be improved over time, a fixed mindset is the belief that

these traits are static and cannot be changed (Yeager & Dweck, 2012). Recent research with undergraduate biology majors indicates that some students possess positive beliefs about making errors during learning (i.e., we benefit from making errors during the learning process), but others possess negative beliefs about such errors (i.e., errors are harmful and to be avoided). Although a growth mindset might be necessary for students in the health professions to succeed academically, it is likely not enough.

A further implication of research on error avoidance (Chouvalova et al., 2024; see also Dresel et al., 2013; Tulis et al., 2017) is that students might be hesitant to use practice questions or other forms of retrieval during study because they worry that being unable to retrieve information or answer questions correctly will signal that they do not know the information, or might not ever be able to learn it (a form of fixed mindset). Such concerns are likely related to students' experience of "imposter syndrome", a pattern of behavior in which people doubt their abilities or fear being exposed as a fraud, even if they have adequate abilities in a domain (such as medicine; see Mullangi & Jagsi, 2019). Students who are high on imposter syndrome are less likely to ask for help and do not want to be "outed" as an imposter (Chen & Son, 2024). Related, many medical students experience forms of academic shame around poor performance and other metrics which suggest they are not performing as well as their peers (and leads to reduced help-seeking behaviors; see Coudret, 2020). Students of the health professions who experience imposter syndrome or feelings of academic shame might avoid completing practice questions or other forms of retrieval practice to avoid such feelings or to avoid being "outed" as an imposter, sacrificing the benefits of retrieval practice to spare their egos.

Other than Coudret (2020), who conducted interviews with medical students about the sources and consequences of their feelings of academic shame, the empirical work above regarding negative beliefs about errors (e.g., Chouvalova et al., 2024) and imposter syndrome (e.g., Chen & Son, 2024) mostly involves undergraduate students. Although we would expect that the same patterns might occur with students of the health professions (i.e., imposter syndrome seems common in medical students; see Mullangi & Jagsi, 2019), these topics need exploration with these specific populations. We know of no research that has examined error avoidance per se in the health professions.

4.2. Addressing Barriers to the Use of Retrieval Practice

Despite the barriers that may deter health profession students from using retrieval practice, strategies exist to increase its adoption. Two recent reviews (Carpenter, 2023; Rivers, 2021) offer recommendations for promoting its use in students broadly.

4.2.1. Improving Knowledge About Retrieval Practice

Educating students about retrieval practice and its benefits could increase students' use of retrieval practice. Table 2 lists resources aimed at informing health profession students, faculty and tutors. However, merely providing information does not usually lead to significant behavioral change (Ariel & Karpicke, 2018; Carpenter, 2023; Morehead et al., 2016; Simone et al., 2023). Rivers (2021) acknowledged these challenges but also highlighted studies demonstrating benefits from explicit instruction (e.g., Einstein et al., 2012; Lineweaver et al., 2019; McCabe, 2011). Health profession students frequently use retrieval practice but often value peer advice over faculty guidance (Ronner & Linkowski, 2020). Thus, engaging senior students to promote its use among peers can be effective (Jodrell, 2010). Additionally, educating students on efficient study techniques can enhance their role as peer learning specialists or tutors. Creating a supportive environment with faculty and administrators may further encourage adoption (Zumbrunn et al., 2014).

4.2.2. Improving Experiences with Retrieval Practice

Another strategy involves providing experience with retrieval practice. Students often revert to less effective methods despite prior use (Hui et al., 2021, 2022; Kirk-Johnson et al., 2019). Because retrieval practice feels difficult, students may mistakenly perceive other more fluent methods as being more effective (Kelley & Lindsay, 1993; Koriat, 2008). Carpenter (2023) emphasized the importance of teaching students that effortful retrieval is beneficial, a form of "desirable difficulties" (Bjork & Bjork, 2023; Carpenter et al., 2020a, 2020b). Explicit feedback can further help students recognize the effectiveness of retrieval practice, increasing subsequent use (Carpenter et al., 2017; Hui et al., 2021, 2022).

Educational technology can also support retrieval practice (Carpenter, 2023). Many digital tools in the health professions offer practice-test questions, performance feedback and analytics that highlight weaknesses. Such tools not only educate students but also reinforce the use of retrieval practice through experience and feedback (cf. Table 2).

4.2.3. Addressing Negative Beliefs About Testing and Errors

Chouvalova et al.'s (2024) research with undergraduate biology majors indicates that, although having a fixed mindset did not relate to students' study strategy selection or use (presumably because they did not think that trying different strategies would matter for their learning), the effect of a growth mindset on students' achievement was mediated by their beliefs about errors. Specifically, holding positive beliefs about errors led students to choose more effective study strategies, such as spaced study and retrieval. As discussed above, more research is needed to better understand how error avoidance and imposter syndrome affect study strategy choices for health profession students, and to develop interventions to address those negative beliefs. Potentially effective interventions might target students' negative beliefs about errors (Chouvalova et al., 2024) and increase their understanding of the positive benefits of effortful processing and retrieval (Carpenter, 2023). Importantly, at least one study has found that the benefits of retrieval practice for learning might be larger for students with a low mastery orientation versus a high mastery orientation (Endres & Eitel, 2024, Experiment 1). As well, the effects can be enhanced with explicit rewards that incentivize learning (Endres & Eitel, 2024, Experiment 2).

4.2.4. Improved Self-Regulated Learning Strategies

Students often choose less-effective strategies for self-regulated learning, even when they are aware of the benefits of retrieval practice (Corazza et al., 2023; Klein et al., 2023; Sheehy et al., 2024). More research is needed to better understand how health profession students engage in retrieval practice during self-regulated learning and how to improve the use of retrieval practice during self-regulated learning. Combining interventions may be necessary for widespread adoption. Several academic programs centered on retrieval practice, such as The Study Smart Program (Biwer et al., 2023), The Start and Stick to Desirable Difficulties framework (de Bruin et al., 2023) and The Knowledge, Belief, Commitment, and Planning framework (McDaniel & Einstein, 2020), all integrate multiple strategies. Of note, both The Study Smart Program and Desirable Difficulties framework have successfully improved performance in health profession students (Biwer et al., 2023).

4.3. Enhancing the Efficacy of Retrieval Practice

There are many opportunities for further improving retention of health professions material across curricula, and arguably it is critical within the health professions where students are preparing for large comprehensive or licensing exams and—arguably—these students' "true final exam" is their clinical practice. Within courses, however, students should be provided with many low-stakes retrieval attempts that can be used formatively

to improve learning in addition to improving metacognition and realization of the utility of retrieval practice. In some universities, formative assessments are implemented (e.g., Arja et al., 2018; Lakhtakia et al., 2022), but the degree to which these assessments are used as retrieval practice (as opposed to providing students with metacognitive awareness or to guide future study) is unclear. In professions where students are being prepared for patient care, using evidence-based learning strategies becomes a moral imperative.

When students first enter the curriculum in the health professions, they have a wide variety of prior knowledge. Because retrieval practice is more effective when retrieval is successful (Chan et al., 2024), some students (i.e., those who understand a topic less are less likely to be able to retrieve information) may need more time with explicit instruction and content review before they engage in retrieval practice (Kalyuga, 2007). Furthermore, students may start engaging with retrieval practice at a factual level to increase retention and understanding. However, health professions education, by its nature, necessitates a considerable amount of transfer of knowledge from the classroom to applied settings (Patel & Cranton, 1983; see Kaminske et al., 2020 for discussion). Often, exams within courses involve case vignettes or patient confederates to which students must apply their factual understanding. While students certainly need to acquire a vast amount of factual knowledge to be successful, utilizing retrieval practice solely for low-level recall of facts may not lead to large performance gains on applied exams. Although data indicates that such flashcard use is positively related to performance on such exams (e.g., Deng et al., 2015; Wu et al., 2021), it is incorrect to assume that flashcards alone can produce sufficient learning or comprehension for excelling on these exams. This is particularly true when students utilize flashcard-based programs where they repeatedly see the same factual items spaced out over time. The effectiveness of these flashcard systems depends on how students utilize them. Often, students will only keep a card in a deck until they have answered it correctly once, dropping it from further retrieval attempts, which is far less effective than completing multiple retrievals (Karpicke & Roediger, 2007). Students might also become so familiar with their deck of cards that they stop engaging in full retrieval of the answers, which is typical of covert retrieval (Sumeracki & Castillo, 2022).

One of the major benefits of flashcard software programs over physical flashcards is that they are often designed to create expanding, spaced retrieval practice. However, it is unclear exactly how the algorithms for many of these programs are designed, whether they are designed to maximize learning, and whether students use them the way they are intended. A notable exception is Anki, an open source digital flashcard program used by students in various disciplines to study for exams, including students in the health professions (Deng et al., 2015). Anki's documentation includes discussion of the benefits of both retrieval practice and expanding retrieval practice (Anki, n.d.). Both factors are key to its design and efficacy. Early algorithms used in Anki software relied on seeing new information frequently and then, based on successful retrieval, expanding intervals (e.g., every three days, every week, every two weeks). This model relied on students feeding information into the algorithm through rating the ease in which they were able to recall information. As noted earlier, because students' metacognition is often misaligned with their performance, less effective retrieval schedules may often be implemented. In addition, when students build up many new cards, it limits their ability to continue the repetition of older cards even if they have not yet mastered that content. For example, a popular Anki card deck designed for the USMLE Step 1 exam, "ANKING", has over 30,000 cards that some students try to master in an 18 month to 2-year period. Attempting to master so many flashcards is extremely time-consuming and can limit students' ability to engage in more effective study behaviors including attending lectures and answering vignette-style practice questions. Newer algorithms have emerged for Anki, including the Free Spaced

Repetition Scheduler algorithm that uses machine learning to learn the user's memory patterns and allows greater options for delaying or speeding up reviews (Ye et al., 2022). Examining ideal schedules and algorithms for such programs is a much-needed area of further research and refinement.

We should note that most of our discussion about the efficacy of retrieval practice has focused on student self-directed learning. Retrieval practice, however, can and should be incorporated into health professions courses. Indeed, institutions have incorporated retrieval practice to varying degrees. In one study, Azzam and Easteal (2021) found that implementing retrieval practice at the beginning of class significantly improved final exam performance in Gross Anatomy. As suggested in Table 2, many university webpages discuss the use of retrieval practice as a curricular tool. However, it is unclear both from the published literature and from these webpages the degree to which retrieval practice is being implemented within and across health professions curriculum and this content likely varies considerably across instructors and institutions.

Admittedly, our review has primarily focused on forms of retrieval practice that involved the recall of explicit information from semantic or episodic memory (e.g., using flashcard software or completing practice questions during study), usually in the service of preparing to answer test questions that also test explicit semantic or episodic memory (e.g., multiple-choice medical licensing exam questions). But either retrieval practice or the final test could involve tests of procedural learning, which reflects our implicit level of skill acquisition (e.g., a medical student's ability to physically insert a urinary catheter, rather than their explicit knowledge of how to do so). To date, however, very few studies in the domains of cognitive psychology or educational psychology have examined the effects of retrieval practice of any kind on skills learning, or of procedural practice on explicit or implicit learning. Even fewer have done so in the context of health professions, but those studies have found positive results. For example, Kromann et al. (2009, 2010) found that explicitly testing a group of medical students on their resuscitation skills just after completing a training course led to higher performance on a final test, compared to a group who completed the same course content but without the intervening test. Although the final test was explicit in nature, the retrieval practice involved implicit recall (performing the task). Conversely, Ndoja et al. (2022) found that explicitly retrieving procedural information–specifically, writing down the steps of an orthopedic surgery–reduced forgetting from pretest to posttest compared to a group that restudied the steps. In this case, although the retrieval practice involved the explicit retrieval (writing down) of the steps involved in the surgery, the final test involved students actually performing the procedure. Such findings indicate that the learning of procedural information in the context of the health professions can benefit from retrieval practice (whether explicit or implicit in nature), and that procedural practice can have similar benefits for learning as does explicit retrieval practice. To date, however, these topics are under-studied in these contexts despite the importance of procedural skills for many aspects of the health professions. An exception has long been surgical procedures, where learning is often done by repetitive practice or simulation (cf. Anton et al., 2019; Rosenthal et al., 2010; Scott et al., 2011). That said, whether or not practicing procedural skills involves "retrieval" is a debate that extends far beyond the boundaries of this review. It is clear that people can improve their skills with practice, but explicit retrieval does not need to occur for this to happen, as people without episodic memory can improve their procedural learning with practice (Squire, 2009).

5. Summary and Conclusions

In this review, we have attempted to summarize and describe the major impact that research on retrieval practice has had on education in the health professions. Research on retrieval practice within cognitive psychology experienced a resurgence in the early 2000s (i.e., Roediger & Karpicke, 2006a) and quickly found its way into medical education (e.g., Larsen et al., 2008, 2009) and the health professions more generally. From there, health profession researchers and educators were quick to embrace retrieval practice, with most empirical studies of the topic in these contexts occurring within the last decade or so. It is difficult, however, to assess the full extent to which retrieval practice is used within health professions education for at least two reasons. First, learners and researchers have likely been using and studying retrieval practice in various forms, using different labels (e.g., distributed or spaced practice, "pimping", etc.). Second, health profession students and trainees receive advice and instruction about learning strategies from several different sources including formal curriculum, MELS, third-party resources and—increasingly—online recommendations from peers (e.g., Reddit, TikTok and YouTube). Thus, while there appears to be a broad adoption of retrieval practice in health professions education, the advice and nuance on how to use retrieval practice is somewhat fractured.

In many ways, the health professions are an ideal setting for adopting retrieval practice. Students and trainees must learn an ever-increasing amount of information, take high-stakes licensing examinations, are likely high-performing students, their training involves a mix of lecture-based instruction and hands-on clinical experience, and involves heavy use of third-party online resources that rely strongly on question-banks. All of these factors contribute to an environment and culture where students are highly incentivized to use flashcards and question banks to make their learning as efficient as possible. Despite the widespread use of retrieval practice, there are still barriers to students using it effectively. First, students may be unaware of the benefits of retrieval practice as many seek information about how to study through unvetted social media sources. Second, students' previous experiences with retrieval practice may dissuade them from further implementing retrieval practice. They may have either used low-effort forms of retrieval practice and been unimpressed with the results (e.g., overly relying on flashcards), or were discouraged by experiencing low levels of fluency during retrieval practice (see Karpicke, 2009). Finally, negative beliefs about errors could lead students to avoid using retrieval practice altogether (Chouvalova et al., 2024).

Retrieval practice has clear benefits for health professions education where the transfer of learning to practice is both clear and important, yet there remains limited research to date on how students use retrieval practice for self-regulated learning. While institutions have increased their use of active learning, retrieval practice as an explicitly used learning tool is still limited. Despite the broad awareness and interest in retrieval practice, it remains under-researched in health professions education. In addition, we could make greater efforts to train faculty and staff in the health professions to allow them to better understand how and why retrieval practice benefits learning. While students are the primary users of retrieval-based strategies, instructors and academic support staff play a crucial role in shaping educational environments and reinforcing effective learning techniques. Many faculty members may be unaware of the robust evidence supporting retrieval practice or how to implement it effectively in both classroom and clinical settings. Providing targeted professional development and training could help close this gap, ensuring that educators are well-equipped to guide students in applying retrieval-based strategies in more purposeful and informed ways.

Author Contributions: Conceptualization, M.J.S., A.N.K., C.N. and K.M.C.; writing—original draft preparation, M.J.S., A.N.K., C.N. and K.M.C.; writing—review and editing, M.J.S., A.N.K., C.N. and K.M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LASSI	Learning and Study Strategies Inventory
LCME	Liaison Committee on Medical Education
MCAT	Medical College Admission Test
MELS	Medical Education Learning Specialist(s)
NBME	National Board of Medical Examiners
PBL	Problem-Based Learning
USMLE	United States Medical Licensing Examination

References

Abbott, E. E. (1909). On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements*, *11*, 159–177. [CrossRef]

Abou-Hanna, J. J., Owens, S. T., Kinnucan, J. A., Mian, S. I., & Kolars, J. C. (2021). Resuscitating the Socratic method: Student and faculty perspectives on posing probing questions during clinical teaching. *Academic Medicine*, *96*, 113–117. [CrossRef]

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*, 659–701. [CrossRef]

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861–876. [CrossRef]

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*, *33*, 1409–1453. [CrossRef]

Ahmed, O. M., Juthani, P. V., Green, M. L., & Moeller, J. J. (2021). Optimizing preclinical learning with retrieval practice: A call to action. *Medical Teacher*, *43*, 718–720. [CrossRef] [PubMed]

Albanese, M. A., & Mitchell, S. (1993). Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic Medicine*, *68*, 52–81. [CrossRef] [PubMed]

Alishahedani, M. E., Sarosi, G. A., & Taylor, J. A. (2019). Implementing survey-based changes and analyzing usage in an online curriculum for surgical interns. *Journal of Surgical Research*, *242*, 87–93. [CrossRef] [PubMed]

AMBOSS. (n.d.). *Internal medicine (ABIM)*. Available online: https://www.amboss.com/us/board-review/internal-medicine (accessed on 25 January 2025).

American Psychological Association. (2018, April 19). *APA dictionary of psychology: "Learning"*. Available online: https://dictionary.apa.org/learning (accessed on 24 June 2025).

Anderson, J. (2013). Can "pimping" kill? The potential effect of disrespectful behavior on patient safety. *JAAPA*, *26*, 53–56. [CrossRef]

Anki. (n.d.). *Anki manual: Background*. Available online: https://docs.ankiweb.net/background.html (accessed on 25 January 2025).

Anton, N. E., Mizota, T., Timsina, L. R., Whiteside, J. A., Myers, E. M., & Stefanidis, D. (2019). Attentional selectivity, automaticity, and self-efficacy predict simulator-acquired skill transfer to the clinical environment. *The American Journal of Surgery*, *217*, 266–271. [CrossRef]

Antonoff, M. B., & D'Cunha, J. (2011, June). Retrieval practice as a means of primary learning: Socrates had the right idea. In *Seminars in thoracic and cardiovascular surgery* (Vol. 23, No. 2, pp. 89–90). WB Saunders. [CrossRef]

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied*, *24*, 43–56. [CrossRef]

Arja, S. B., Acharya, Y., Alezaireg, S., Ilavarasan, V., Ala, S., & Arja, S. B. (2018). Implementation of formative assessment and its effectiveness in undergraduate medical education: An experience at a Caribbean Medical School. *MedEdPublish*, *7*, 131. [CrossRef] [PubMed]

Artino, A. R., Brydges, R., & Gruppen, L. D. (2015). *Self-regulated learning in healthcare profession education: Theoretical perspectives and research methods* (1st ed.). JohnWiley & Sons, Ltd.

Augustin, M. (2014). How to learn effectively in medical school: Test yourself, learn actively, and repeat in intervals. *The Yale Journal of Biology and Medicine*, *87*, 207–212. [PubMed]

Azzam, M. B., & Easteal, R. A. (2021). Retrieval practice for improving long-term retention in anatomical education: A quasi-experimental study. *Medical Science Educator*, *31*, 1305–1310. [CrossRef] [PubMed]

Bagley, S. (2020). The flipped classroom, lethal mutations, and the didactical contract: A cautionary tale. *Primus*, *30*, 243–260. [CrossRef]

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, *128*, 612–637. [CrossRef]

Barrison, P. D., Balczewski, E. A., Capellari, E., Landis-Lewis, Z., & Vinson, A. H. (2024). Electronic flashcards in health professions education: A scoping review. *Academic Medicine*, *10*, 1097. [CrossRef]

Barron, A. B., Hebets, E. A., Cleland, T. A., Fitzpatrick, C. L., Hauber, M. E., & Stevens, J. R. (2015). Embracing multiple definitions of learning. *Trends in Neurosciences*, *38*, 405–407. [CrossRef]

Barrows, H. S. (1986). A taxonomy of problem-based learning methods. *Medical Education*, *20*, 481–486. [CrossRef]

Barry, E. S., Merkebu, J., & Varpio, L. (2022). Understanding state-of-the-art literature reviews. *Journal of Graduate Medical Education*, *14*, 659–662. [CrossRef]

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*, 55–68. [CrossRef]

Biwer, F., de Bruin, A., & Persky, A. (2023). Study smart—Impact of a learning strategy training on students' study behavior and academic performance. *Advances in Health Sciences Education*, *28*, 147–167. [CrossRef] [PubMed]

Bjork, E. L., & Bjork, R. A. (2023). Introducing desirable difficulties into practice and instruction: Obstacles and opportunities. In C. E. Overson, C. M. Hakala, L. L. Kordonowy, & V. A. Benassi (Eds.), *In their own words: What scholars and teachers want you to know about why and how to apply the science of learning in your academic setting* (pp. 19–30). Society for the Teaching of Psychology.

Blake, R. L., Hosokawa, M. C., & Riley, S. L. (2000). Student performances on Step 1 and Step 2 of the United States Medical Licensing Examination following implementation of a problem-based learning curriculum. *Academic Medicine*, *75*, 66–70. [CrossRef] [PubMed]

Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: Comparing intended versus actual study behaviour. *Memory*, *25*, 784–792. [CrossRef] [PubMed]

Brancati, F. L. (1989). The art of pimping. *JAMA*, *262*, 89–90. [CrossRef]

Briñol, P., Petty, R. E., & Tormala, Z. L. (2006). The malleable meaning of subjective ease. *Psychological Science*, *17*, 200–206. [CrossRef]

Buchin, Z. L., & Mulligan, N. W. (2023). Retrieval-based learning and prior knowledge. *Journal of Educational Psychology*, *115*, 22–35. [CrossRef]

Carpenter, S. K. (2023). Encouraging students to use retrieval practice: A review of emerging research from five types of interventions. *Educational Psychology Review*, *35*, 96. [CrossRef]

Carpenter, S. K., Endres, T., & Hui, L. (2020a). Students' use of retrieval in self-regulated learning: Implications for monitoring and regulating effortful learning experiences. *Educational Psychology Review*, *32*, 1029–1054. [CrossRef]

Carpenter, S. K., Rahman, S., Lund, T. J. S., Armstrong, P. I., Lamm, M. H., Reason, R. D., & Cofman, C. R. (2017). Students' use of optional online reviews and its relationship to summative assessment outcomes in introductory biology. *CBE Life Sciences Education*, *16*, ar23. [CrossRef]

Carpenter, S. K., Witherby, A. E., & Tauber, S. K. (2020b). On students' (mis)judgments of learning and teaching effectiveness. *Journal of Applied Research in Memory & Cognition*, *9*, 137–151. [CrossRef]

Chan, J. C. K., Davis, S. D., Yurtsever, A., & Myers, S. J. (2024). The magnitude of the testing effect is independent of retrieval practice performance. *Journal of Experimental Psychology: General*, *153*, 1816–1837. [CrossRef] [PubMed]

Chen, S., & Son, L. K. (2024). High Impostors are more hesitant to ask for help. *Behavioral Sciences*, *14*, 810. [CrossRef] [PubMed]

Chouvalova, A., Navlekar, A. S., Mills, D. J., Adams, M., Daye, S., De Anda, F., & Limeri, L. B. (2024). Undergraduates' reactions to errors mediates the association between growth mindset and study strategies. *International Journal of STEM Education*, *11*, 26. [CrossRef]

Cogliano, M., Kardash, C. M., & Bernacki, M. L. (2019). The effects of retrieval practice and prior topic knowledge on test performance and confidence judgments. *Contemporary Educational Psychology*, *56*, 117–129. [CrossRef]

Coppola, K. M., Rashid, H., Terregino, C. A., & Lebeau, R. (2023). Determining root causes of poor academic performance to provide wraparound support for preclerkship medical students. *Academic Medicine*, *10*, 1097. [CrossRef]

Corazza, L., Shirkhani, S., Berberat, P. O., & Wijnen-Meijer, M. (2023). Structured interviews on self-regulated learning strategies of medical students in the final year of medical school. *BMC Medical Education*, *23*, 604. [CrossRef]

Coudret, D. J. (2020). *Turning the spotlight on shame: Fostering adaptive responses to feelings of academic shame in medical students* [Doctoral dissertation, Arizona State University].

Coumarbatch, J., Robinson, L., Thomas, R., & Bridge, P. D. (2010). Strategies for identifying students at risk for USMLE Step 1 failure. *Family Medicine*, *42*, 105–110.

Cutting, M. F., & Saks, N. S. (2012). Twelve tips for utilizing principles of learning to support medical education. *Medical Teacher*, *34*, 20–24. [CrossRef]

de Bruin, A. B. H., Biwer, F., Hui, L., Onan, E., David, L., & Wiradhany, W. (2023). Worth the effort: The start and stick to desirable difficulties (S2D2) framework. *Educational Psychology Review*, *35*, 41. [CrossRef]

de Lima, M. F. R., & Buratto, L. G. (2024). Retrieval practice effect and individual differences: Current status and future directions. *Journal of Cognitive Psychology*, *36*, 443–456. [CrossRef]

Deng, F., Gluckstein, J. A., & Larsen, D. P. (2015). Student-directed retrieval practice is a predictor of medical licensing examination performance. *Perspectives on Medical Education*, *4*, 308–313. [CrossRef]

Detsky, A. S. (2009). The art of pimping. *JAMA*, *301*, 1379–1381. [CrossRef]

Didarloo, A., & Khalkhali, H. R. (2014). Assessing study skills among a sample of university students: An Iranian survey. *Journal of Educational Evaluation for Health Professions*, *11*, 1148910. [CrossRef]

Dobson, J., Linderholm, T., & Perez, J. (2018). Retrieval practice enhances the ability to evaluate complex physiology information. *Medical Education*, *52*, 513–525. [CrossRef] [PubMed]

Dobson, J. L., Linderholm, T., & Stroud, L. (2019). Retrieval practice and judgements of learning enhance transfer of physiology information. *Advances in Health Sciences Education*, *24*, 525–537. [CrossRef] [PubMed]

Donker, S. C., Vorstenbosch, M. A., Gerhardus, M. J., & Thijssen, D. H. (2022). Retrieval practice and spaced learning: Preventing loss of knowledge in Dutch medical sciences students in an ecologically valid setting. *BMC Medical Education*, *22*, 65. [CrossRef] [PubMed]

Dresel, M., Schober, B., Ziegler, A., Grassinger, R., & Steuer, G. (2013). Affective-motivational adaptivity and action adaptivity reactions to errors during learning. *Zeitschrift für pädagogische Psychologie*, *27*, 255–271. [CrossRef]

Einstein, G. O., Mullet, H. G., & Harrison, T. L. (2012). The testing effect: Illustrating a fundamental concept and changing study strategies. *Teaching of Psychology*, *39*, 190–193. [CrossRef]

Endres, T., & Eitel, A. (2024). Motivation brought to the test: Successful retrieval practice is modulated by mastery goal orientation and external rewards. *Applied Cognitive Psychology*, *38*, e4160. [CrossRef]

England, B. D., & Serra, M. J. (2012). The contributions of anchoring and past-test performance to the underconfidence-with-practice effect. *Psychonomic Bulletin & Review*, *19*, 715–722. [CrossRef]

Ezeala, C. C., & Siyanga, N. (2015). Analysis of the study skills of undergraduate pharmacy students of the University of Zambia School of Medicine. *Journal of Educational Evaluation for Health Professions*, *12*. [CrossRef] [PubMed]

Friedlander, M. J., Andrews, L., Armstrong, E. G., Aschenbrenner, C., Kass, J. S., Ogden, P., Schwartzstein, R., & Viggiano, T. R. (2011). What can medical education learn from the neurobiology of learning? *Academic Medicine*, *86*, 415–420. [CrossRef] [PubMed]

Furness, T., Wilson, C., Ryan, A., & Judd, T. (2024). The attitudes towards learning and study habits of Australian medical students. *Medical Teacher*, *47*, 1199–1210. [CrossRef]

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*, Serial No. 40.

Gebremedhin, D. (2012). *The Step 1 Method: A Step by Step Guide to Success on the USMLE Step 1 Exam* (1st ed.). lulu.com.

Giuliodori, M. J., Lujan, H. L., & DiCarlo, S. E. (2008). Collaborative group testing benefits high-and low-performing students. *Advances in Physiology Education*, *32*, 274–278. [CrossRef]

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399. [CrossRef]

Goldman, M., Bryan, J., & Lucke-Wold, B. (2024). Evidence-based educational algorithm "Anki" for optimization of medical education. *Journal of Biomed Research*, *5*, 1–7. Available online: https://www.researchgate.net/publication/383200347 (accessed on 12 February 2025). [CrossRef]

Gooding, H. C., Mann, K., & Armstrong, E. (2017). Twelve tips for applying the science of learning to health professions education. *Medical Teacher*, *39*, 26–31. [CrossRef]

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, *26*, 91–108. [CrossRef]

Green, M. L., Moeller, J. J., & Spak, J. M. (2018). Test-enhanced learning in health professions education: A systematic review: BEME Guide No. 48. *Medical Teacher*, *40*, 337–350. [CrossRef] [PubMed]

Halamish, V. (2018). Pre-service and in-service teachers' metacognitive knowledge of learning strategies. *Frontiers in Psychology*, *9*, 2152. [CrossRef] [PubMed]

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*, 126–134. [CrossRef]

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1191–1206. [CrossRef]

Hui, L., de Bruin, A. B. H., Donkers, J., & van Merriënboer, J. J. G. (2021). Does individual performance feedback increase the use of retrieval practice? *Educational Psychology Review*, *33*, 1835–1857. [CrossRef]

Hui, L., de Bruin, A. B. H., Donkers, J., & van Merriënboer, J. J. G. (2022). Why students do (or do not) choose retrieval practice: Their perceptions of mental effort during task performance matter. *Applied Cognitive Psychology*, *36*, 433–444. [CrossRef]

Jimenez, M. J. D., Kantak, P., Raskin, J., & Kantak, P. A. (2024). Why Pimping Works: The Neurophysiology of Emotional Memories. *Cureus*, *16*, e64237. [CrossRef]

Jodrell, D. (2010). Social-Identity and Self-Efficacy Concern for Disability Labels. *Psychology Teaching Review*, *16*, 111–121. [CrossRef]

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*, 509–539. [CrossRef]

Kaminske, A. N., Kuepper-Tetzel, C. E., Nebel, C. L., Sumeracki, M. A., & Ryan, S. P. (2020). Transfer: A review for biology and the life sciences. *CBE—Life Sciences Education*, *19*, es9. [CrossRef]

Kann, M. R., Chen, D., Huang, G. W., Pugazenthi, S., & Zehnder, N. G. (2024). A medical student online community: A qualitative analysis of step 1 subreddit posts. *Medical Science Educator*, *35*, 425–436. [CrossRef] [PubMed]

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*, 469–486. [CrossRef] [PubMed]

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*, 157–163. [CrossRef]

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. T. Wixted (Ed.), *Cognitive psychology of memory, of learning and memory: A comprehensive reference* (Vol. 2, pp. 487–514). Academic Press.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*, 317–326. [CrossRef]

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory*, *17*, 471–479. [CrossRef]

Karpicke, J. D., & Roediger, H. L., III. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162. [CrossRef]

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, *32*, 1–24. [CrossRef]

Khalil, M. K., Hawkins, H. G., Crespo, L. M., & Buggy, J. (2017). The relationship between Learning and Study Strategies Inventory (LASSI) and academic performance in medical schools. *Medical Science Educator*, *27*, 315–320. [CrossRef]

Khalil, M. K., Williams, S. E., & Gregory Hawkins, H. (2018). Learning and study strategies correlate with medical students' performance in anatomical sciences. *Anatomical Sciences Education*, *11*, 236–242. [CrossRef]

Khalil, M. K., Williams, S. E., & Hawkins, H. G. (2020). The use of Learning and Study Strategies Inventory (LASSI) to investigate differences between low vs high academically performing medical students. *Medical Science Educator*, *30*, 287–292. [CrossRef]

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, *115*, 1–31. [CrossRef]

Klein, H., Allen, E., & McCarthy, S. (2023). Characterizing the successful student in virtual gross anatomy. *American Journal of Science Education Research*, *AJSER-115*, 61–74.

Kleinsorgen, C., von Köckritz-Blickwede, M., Naim, H. Y., Branitzki-Heinemann, K., Kankofer, M., Mándoki, M., Adler, M., Tipold, A., & Ehlers, J. P. (2018). Impact of virtual patients as optional learning material in veterinary biochemistry education. *Journal of Veterinary Medical Education*, *45*, 177–187. [CrossRef]

Kolomitro, K., MacKenzie, L. W., Lockridge, M., & Clohosey, D. (2020). Problem-solving strategies used in anatomical multiple-choice questions. *Health Science Reports*, *3*, 1–7. [CrossRef]

Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition*, *36*, 416–428. [CrossRef] [PubMed]

Koriat, A., & Ackerman, R. (2010). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science*, *13*, 441–453. [CrossRef] [PubMed]

Kriechbaum, V. M., & Bäuml, K.-H. T. (2024). Retrieval practice can promote new learning with both related and unrelated prose materials. *Journal of Applied Research in Memory and Cognition*, *13*, 319–328. [CrossRef]

Kromann, C. B., Bohnstedt, C., Jensen, M. L., & Ringsted, C. (2010). The testing effect on skills learning might last 6 months. *Advances in Health Sciences Education*, *15*, 395–401. [CrossRef]

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education*, *43*, 21–27. [CrossRef]

Lakhtakia, R., Otaki, F., Alsuwaidi, L., & Zary, N. (2022). Assessment as learning in medical education: Feasibility and perceived impact of student-generated formative assessments. *JMIR Medical Education*, *8*, e35820. [CrossRef]

Larsen, D. P., Butler, A. C., Aung, W. Y., Corboy, J. R., Friedman, D. I., & Sperling, M. R. (2015). The effects of test-enhanced learning on long-term retention in AAN annual meeting courses. *Neurology*, *84*, 748–754. [CrossRef]

Larsen, D. P., Butler, A. C., Lawson, A. L., & Roediger, H. L. (2013). The importance of seeing the patient: Test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Advances in Health Sciences Education*, *18*, 409–425. [CrossRef] [PubMed]

Larsen, D. P., Butler, A. C., & Roediger, H. L., III. (2008). Test-enhanced learning in medical education. *Medical Education*, *42*, 959–966. [CrossRef] [PubMed]

Larsen, D. P., Butler, A. C., & Roediger, H. L., III. (2009). Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial. *Medical Education*, *43*, 1174–1181. [CrossRef] [PubMed]

Lawrence, E. C., Dine, C. J., & Kogan, J. R. (2023). Preclerkship medical students' use of third-party learning resources. *JAMA Network Open*, *6*, e2345971. [CrossRef]

Lecturio. (2024, May 6). *Active Recall and retrieval practice in medical education*. Available online: https://www.lecturio.com/inst/pulse/retrieval-based-learning-strategies-in-medical-education/ (accessed on 25 January 2025).

Lineweaver, T. T., Hall, A. C., Hilycord, D., & Vitelli, S. E. (2019). Introducing and evaluating a "Study Smarter, Not Harder" study tips presentation offered to incoming students at a four-year university. *Journal of the Scholarship of Teaching and Learning*, *19*, 16–46. [CrossRef]

Little, J. L., Frickey, E. A., & Fung, A. K. (2019). The role of retrieval in answering multiple-choice questions. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *45*, 1473–1485. [CrossRef]

Logan, N. S., Hatch, R. A., & Logan, H. L. (1975). Massed vs. distributed practice in learning dental psycho-motor skills. *Journal of Dental Education*, *39*, 87–91. [CrossRef]

Loma Linda School of Medicine. (2021, July 7). *The science (and illusions) of learning: Instructional design tips for teachers to enhance learners' long-term memory retention*. Available online: https://medicine.llu.edu/sites/medicine.llu.edu/files/2021-07/Handout-The-Sciences-Illusions-for-Learning-for-the-Long-Term.pdf (accessed on 25 January 2025).

Matsko, C., & Cervantes, J. (2024). Third-party resources: The new wave of medical education. *Medical Science Educator*, *35*, 1089–1092. [CrossRef]

McAndrew, M., Kamboj, R. S., & Pierre, G. C. (2015). Do dental students use optimal study strategies? *Journal of Dental Education*, *79*, 33–37. [CrossRef]

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*, 462–476. [CrossRef]

McCarthy, C. P., & McEvoy, J. W. (2015). Pimping in medical education: Lacking evidence and under threat. *JAMA*, *314*, 2347–2348. [CrossRef]

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494–513. [CrossRef]

McDaniel, M. A., & Einstein, G. O. (2020). Training learning strategies to promote self-regulation and transfer: The knowledge, belief, commitment, and planning framework. *Perspectives on Psychological Science*, *15*, 1363–1381. [CrossRef]

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192–201. [CrossRef]

McEvoy, J. W., Shatzer, J. H., Desai, S. V., & Wright, S. M. (2019). Questioning style and pimping in clinical education: A quantitative score derived from a survey of internal medicine teaching faculty. *Teaching and Learning in Medicine*, *31*, 53–64. [CrossRef]

Metz, M. J., Immekus, J. C., Lyle, K. B., & George, C. E. (2021). Effects of active engagement and spaced retrieval practice on knowledge and application of a self-assessment rubric. *Journal of Dental Education*, *85*, 1786–1794. [CrossRef]

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*, 257–271. [CrossRef]

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: A review of applied research. *Frontiers in Education*, *4*. [CrossRef]

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533. [CrossRef]

Mullangi, S., & Jagsi, R. (2019). Imposter syndrome: Treat the cause, not the symptom. *JAMA*, *322*, 403–404. [CrossRef]

Musselman, L. J., MacRae, H. M., Reznick, R. K., & Lingard, L. A. (2005). 'You learn better under the gun': Intimidation and harassment in surgical education. *Medical Education*, *39*, 926–934. [CrossRef]

Nagarur, A., McEvoy, J. W., Hirsh, D. A., & James, B. C. (2019). Words matter: Removing the word pimp from medical education discourse. *The American Journal of Medicine*, *132*, e813–e814. [CrossRef] [PubMed]

Ndoja, S., Dion, C. A., Pirshahid, A. A., Charron, B. P., Durocher, A., McCarton, A., & LeBel, M. E. (2022). Active retrieval improves procedural learning in orthopedic surgery. *Journal of Surgical Education*, *79*, 1308–1314. [CrossRef] [PubMed]

O'Sullivan, S., Hageh, C. A., Dimassi, Z., Alsoud, L. O., Presley, D., & Ibrahim, H. (2024). Exploring challenges in learning and study skills among first-year medical students: A case study. *BMC Medical Education*, *24*, 869. [CrossRef]

Pan, S. C., Cooke, J., Little, J. L., McDaniel, M. A., Foster, E. R., Connor, L. T., & Rickard, T. C. (2019). Online and clicker quizzing on jargon terms enhances definition-focused but not conceptually focused biology exam performance. *CBE—Life Sciences Education*, *18*, ar54. Available online: https://www.lifescied.org/doi/pdf/10.1187/cbe.18-12-0248 (accessed on 27 June 2025). [CrossRef]

Pan, S. C., Dunlosky, J., Xu, K. M., & Ouwehand, K. (2024). Emerging and future directions in test-enhanced learning research. *Educational Psychology Review*, *36*, 20. [CrossRef]

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*, 710–756. [CrossRef]

Pan, S. C., Sana, F., Samani, J., Cooke, J., & Kim, J. A. (2020). Learning from errors: Students' and instructors' practices, attitudes, and beliefs. *Memory*, *28*, 1105–1122. [CrossRef]

Patel, V. L., & Cranton, P. A. (1983). Transfer of student learning in medical education. *Academic Medicine*, *58*, 126–135. [CrossRef] [PubMed]

Rashid, H., Drake, E., Lebeau, R., Meka, J., & Coppola, K. M. (2024, November 8–12). *A national survey of academic support programs across US medical schools*. Learn, Serve, Lead: The AAMC Annual Meeting, Atlanta, GA, USA.

Rawson, K. A. (2015). The status of the testing effect for complex materials: Still a winner. *Educational Psychology Review*, *27*, 327–331. [CrossRef]

Reifler, D. R. (2015). The pedagogy of pimping: Educational rigor or mistreatment? *JAMA*, *314*, 2355–2356. [CrossRef]

Rivers, M. L. (2021). Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review*, *33*, 823–862. [CrossRef]

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology*, *82*, 416–425. [CrossRef] [PubMed]

Roediger, H. L., III, & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249—255. [CrossRef] [PubMed]

Roediger, H. L., III, & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210. [CrossRef]

Rogers, M. E., Creed, P. A., Searle, J., & Nicholls, S. L. (2016). Coping with medical training demands: Thinking of dropping out, or in it for the long haul. *Studies in Higher Education*, *41*, 1715–1732. [CrossRef]

Ronner, L., & Linkowski, L. (2020). Online forums and the "Step 1 Climate": Perspectives from a medical student reddit user. *Academic Medicine*, *95*, 1329–1331. [CrossRef]

Rosenthal, M. E., Ritter, E. M., Goova, M. T., Castellvi, A. O., Tesfay, S. T., Pimentel, E. A., Hartzler, R., & Scott, D. J. (2010). Proficiency-based fundamentals of laparoscopic surgery skills training results in durable performance improvement and a uniform certification pass rate. *Surgical Endoscopy*, *24*, 2453–2457. [CrossRef]

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432–1463. [CrossRef]

Rucker, L., Rucker, G., Nguyen, A., Noel, M., Marroquin, M., Streja, E., & Hennrikus, E. (2023). Medical faculty and medical student opinions on the utility of questions to teach and evaluate in the clinical environment. *Medical Science Educator*, *33*, 669–678. [CrossRef]

Saks, N. S., Zingale, C. M., & Stewart, D. G. (1998). *How to excel in medical school*. J&S Publishing Company, Inc.

Scott, D. J., Pugh, C. M., Ritter, E. M., Jacobs, L. M., Pellegrini, C. A., & Sachdeva, A. K. (2011). New directions in simulation-based surgical education and training: Validation and transfer of surgical skills, use of nonsurgeons as faculty, use of simulation to screen and select surgery residents, and long-term follow-up of learners. *Surgery*, *149*, 735–744. [CrossRef] [PubMed]

Selva-Rodriguez, A., & Sandars, J. (2023). Twelve tips for providing academic remediation to widening access learners in medical education. *Medical Teacher*, *45*, 1112–1117. [CrossRef] [PubMed]

Serra, M. J., & Ariel, R. (2014). People use the memory for past-test heuristic as an explicit cue for judgments of learning. *Memory & Cognition*, *42*, 1260–1272. [CrossRef] [PubMed]

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1258–1266. [CrossRef]

Sheehy, R., Scott, D., Davis, D., Roffler, M., Sweatman, T., & Nemec, E. (2024). Medical student use of practice questions in their studies: A qualitative study. *BMC Medical Education*, *24*, 1181. [CrossRef]

Simone, P. M., Whitfeld, L. C., Bell, M. C., Kher, P., & Tamashiro, T. (2023). Shifting students toward testing: Impact of instruction and context on self-regulated learning. *Cognitive Research: Principles & Implications*, *8*, 14. [CrossRef]

Sisa, I., Garcés, M. S., Crespo-Andrade, C., & Tobar, C. (2023). Improving learning and study strategies in undergraduate medical students: A pre-post study. *Healthcare*, *11*, 375. [CrossRef]

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656. [CrossRef]

Squire, L. R. (2009). The legacy of patient HM for neuroscience. *Neuron*, *61*, 6–9. [CrossRef]

Steed, K. S., & Kadavakollu, S. (2019). How to prepare for the Medical College Admissions Test (MCAT): Six important tips for pre-medical students from rural areas. *Medical Science Educator*, *29*, 1147–1153. [CrossRef]

Stoddard, H. A., & O'Dell, D. V. (2016). Would Socrates have actually used the "Socratic Method" for clinical teaching? *Journal of General Internal Medicine*, *31*, 1092–1096. [CrossRef] [PubMed]

Student Doctor Network. (2018, August 14). *Gain a deeper understanding with the power of test-enhanced learning*. Available online: https://www.studentdoctor.net/2018/08/14/test-enhanced-learning/ (accessed on 25 January 2025).

Sumeracki, M. A., & Castillo, J. (2022). Covert and overt retrieval practice in the classroom. *Translational Issues in Psychological Science*, *8*, 282–293. [CrossRef]

Swan Sein, A., Cuffney, F., & Clinchot, D. (2020). How to help students strategically prepare for the MCAT exam and learn foundational knowledge needed for medical school. *Academic Medicine*, *95*, 484. [CrossRef] [PubMed]

Swan Sein, A., Dathatri, S., & Bates, T. A. (2021a). Twelve tips on guiding preparation for both high-stakes exams and long-term learning. *Medical Teacher*, *43*, 518–523. [CrossRef]

Swan Sein, A., Rashid, H., Meka, J., Amiel, J., & Pluta, W. (2021b). Twelve tips for embedding assessment for and as learning practices in a programmatic assessment system. *Medical Teacher*, *43*, 300–306. [CrossRef]

Thadani, R. A., Swanson, D. B., & Galbraith, R. M. (2000). A preliminary analysis of different approaches to preparing for the USMLE Step 1. *Academic Medicine*, *75*, S40–S42. [CrossRef]

The Robert Larner College of Medicine at University of Vermont. (n.d.). *The science of learning learning theory: Retrieval practice*. Available online: https://www.med.uvm.edu/docs/retrieval_practicecfcf5c6131ef4ab1b990bd156be8b6ea/active_learning/the_science_of_learning-retrieval_practice.pdf (accessed on 25 January 2025).

Thompson, C. P., & Hughes, M. A. (2023). The effectiveness of spaced learning, interleaving, and retrieval practice in radiology education: A systematic review. *Journal of the American College of Radiology*. [CrossRef]

TrueLearn. (n.d.). *Goodbye, Q-banks. Hello, smartbanks*. Available online: https://truelearn.com/smartbanks/ (accessed on 25 January 2025).

Trumble, E., Lodge, J., Mandrusiak, A., & Forbes, R. (2024). Systematic review of distributed practice and retrieval practice in health professions education. *Advances in Health Sciences Education*, *29*, 689–714. [CrossRef]

Tulis, M., Steuer, G., & Dresel, M. (2017). Positive beliefs about errors as an important element of adaptive individual dealing with errors during academic learning. *Educational Psychology*, *38*, 139–158. [CrossRef]

UConn School of Nursing—University of Connecticut. (n.d.). *Learning science strategies series, module 3: Retrieval practice*. Available online: https://nursing.media.uconn.edu/wp-content/uploads/sites/3147/2022/03/3_Retrieval-practice-handout_VanHoof.pdf (accessed on 25 January 2025).

UWorld. (2021, January 19). *Practice questions: More important than ever for NCLEX success*. Available online: https://nursing.uworld.com/blog/practice-questions-more-important-than-ever-for-nclex-success (accessed on 25 January 2025).

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*, 247–264. [CrossRef]

Van Hoof, T. J., Leary, C. M., Banfi, V., & Polifroni, E. C. (2023). A scoping review of retrieval practice (test-enhanced learning) in nursing education. *Nursing Education Perspectives*, *44*, 341–346. [CrossRef]

Walters, J. A., Croen, L. G., Brown Weissman, Z., & Reichgott, M. J. (1999). A small group, problem-based learning approach to preparing students to retake Step 1 of the United States Medical Licensing Examination. *Teaching and Learning in Medicine*, *11*, 85–88. [CrossRef]

Wear, D., Kokinova, M., Keck-McNulty, C., & Aultman, J. (2005). Research basic to medical education: Pimping: Perspectives of 4th year medical students. *Teaching and Learning in Medicine*, *17*, 184–191. [CrossRef] [PubMed]

Weidman, J., & Baker, K. (2015). The cognitive science of learning: Concepts and strategies for the educator and learner. *Anesthesia & Analgesia*, *121*, 1586–1599. [CrossRef]

Weinstein, C. E., Palmer, D. R., & Schulte, A. C. (1987). *LASSI: Learning and study strategies inventory*. H & H Publishing Company.

Weinstein, C. E., Palmer, D. R., & Schulte, A. C. (2002). *LASSI. User's manual for those administering learning and study strategies inventory*. H & H Publishing Company.

Witherby, A. E., Babineau, A. L., & Tauber, S. K. (2025). Students', Teachers', and Parents' Knowledge About and Perceptions of Learning Strategies. *Behavioral Sciences*, *15*, 160. [CrossRef] [PubMed]

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition*, *3*, 214–221. [CrossRef]

Wu, W., Garcia, K., Chandrahas, S., Siddiqui, A., Baronia, R., & Ibrahim, Y. (2021). Predictors of performance on USMLE step 1. *The Southwest Respiratory and Critical Care Chronicles*, *9*, 63–72. [CrossRef]

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*, 399–435. [CrossRef]

Ye, J., Su, J., & Cao, Y. (2022, August 14–18). *A stochastic shortest path algorithm for optimizing spaced repetition scheduling*. 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 4381–4390), Washington, DC, USA. [CrossRef]

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, *47*, 302–314. [CrossRef]

Yeo, D. J., & Fazio, L. K. (2019). The optimal learning strategy depends on learning goals and processes: Retrieval practice versus worked examples. *Journal of Educational Psychology*, *111*, 73–90. [CrossRef]

Yudcovitch, L., & Hayes, J. R. (2014). Case-based student performance: Socratic method vs. passive presentation. *Optometric Education*, *40*, 37–43.

Zhang, C., Rauchwarger, A., Toth, C., & O'Connell, M. (2005). Student USMLE Step 1 preparation and performance. *Advances in Health Sciences Education*, *9*, 291–297. [CrossRef]

Zimmerman, B. J., & Pons, M. M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Education Research Journal*, *23*, 614–628. [CrossRef]

Zumbrunn, S., McKim, C., Buhs, E., & Hawley, L. R. (2014). Support, belonging, motivation, and engagement in the college classroom: A mixed method study. *Instructional Science*, *42*, 661–684. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

The Role of Answer Justification in Multiple-Choice Testing: Effects on Performance and Metacognitive Accuracy

Spenser A. Clark [1], Michelle L. Rivers [2] and Acacia L. Overono [3,*]

[1] Department of Instructional Technology & Learning Sciences, Utah State University, Logan, UT 84322, USA; spenser.clark@usu.edu
[2] Department of Psychology, Santa Clara University, Santa Clara, CA 95053, USA; mrivers@scu.edu
[3] Psychology and Counseling Department, Utah Valley University, Orem, UT 84058, USA
* Correspondence: acacia.overono@uvu.edu

Abstract

Multiple-choice (MC) tests are widely used in educational settings but have been criticized for promoting passive recognition rather than active retrieval. Our research explores how adding a simple component to MC tests—answer justification—influences test performance and metacognitive accuracy. Across two experiments, university students studied a textbook chapter and completed either a standard MC test (MC-only group) or an MC test requiring them to justify their answers (answer justification group). Participants also provided predictive and postdictive metacognitive judgments. The results showed that the answer justification group significantly outperformed the MC-only group on an immediate test (Experiments 1 and 2) and scored numerically higher on a delayed test two days later (Experiment 2). Further, some initial evidence suggested that metacognitive accuracy was influenced by test type, but future research is needed. These findings support a retrieval-based explanation: generating answer justifications increases test performance by strengthening memory through elaborative retrieval. This study demonstrates that incorporating answer justification into MC tests may improve learning and metacognitive accuracy. We also offer practical suggestions for classroom implementation, considering that answer justification boosts test performance but also imposes a time cost compared to standard MC tests.

Keywords: multiple-choice testing; assessment; metacognition; elaborative retrieval; self-explanation

1. Introduction

What if answering a multiple-choice (MC) question became more than just a test of recognition? Imagine students engaging deeply with each question, weighing each response as a possible solution, and building stronger connections to the material to enhance their understanding. Standard MC questions, while efficient, often fail to achieve this level of cognitive engagement. How might we redesign MC tests to foster meaningful learning?

Our study explores how incorporating answer justification—a process where learners explain their chosen response—can enhance the educational value of MC questions. Drawing on principles of generative learning, we hypothesize that answer justification encourages elaborative retrieval and metacognitive monitoring, prompting learners to connect new information with prior knowledge and critically evaluate their reasoning.

For the remainder of the Introduction, we discuss prior research that informs our specific predictions and outline how we evaluated these predictions across two experiments.

1.1. Improving Multiple-Choice Questions with Answer Justification

A significant challenge in education is to develop teaching and assessment methods that are efficient, adaptable, and effective in promoting meaningful learning. This challenge becomes even more pressing in settings like large classrooms and massive open online courses (MOOCs), where individualized feedback is limited. Multiple-choice (MC) questions are nearly ubiquitous in the classroom due to their ease of grading and scalability. However, they have been criticized for encouraging guesswork, relying on recognition rather than retrieval processes, and failing to measure complex understanding effectively (e.g., Brown & Abdulnabi, 2017; Greving & Richter, 2022; Ozuru et al., 2013). Despite these criticisms, recent research has shown that well-constructed MC tests can foster meaningful retrieval processes and improve subsequent recall of related information, challenging assumptions about their limitations (e.g., J. L. Little et al., 2012, 2019; J. L. Little & Bjork, 2015). For example, research by Sparck et al. (2016) demonstrated that a confidence-weighted MC format—in which test takers indicate their relative confidence among alternative responses—significantly enhanced later recall of relevant information compared to standard MC questions. This finding supports the idea that modifying MC tests to promote active, productive retrieval can offset their inherent reliance on passive recognition and ultimately improve learning outcomes. Another promising enhancement to standard MC questions is *answer justification*, which allows (e.g., Nield & Wintre, 1986) or requires (e.g., Bassett, 2016) learners to explain their reasoning for selecting a particular response at the time of testing.

Initial research on answer justification primarily examined its positive influence on student perceptions of MC questions (e.g., Gorrell & Cramond, 1988; Kottke, 2001; Nield & Wintre, 1986; Dodd & Leal, 1988). However, relatively few studies have explored its impact on student performance. Classroom-based research suggests that incorporating answer justification can enhance retrieval during testing compared to standard MC questions (Koretsky et al., 2016; McKeachie et al., 1955; Overono, in press; Papadopoulos et al., 2022). For example, Papadopoulos et al. (2022) gave students the option to complete in-class MC questions via an audience response system (e.g., "clickers"), with four of the eight questions including an optional justification component (limited to 140 characters). Students provided justifications for 76% of these questions, and the results showed that students earned high scores on questions that included the justification option compared to questions without a justification component. Additionally, similar studies that used MC and answer justification questions in classes (Koretsky et al., 2016; Overono, in press) found small to moderate effect sizes, suggesting an increased likelihood of selecting the correct answer when MC questions included a justification component. These classroom studies provide preliminary evidence that answer justification enhances performance on MC questions, but what explains its effectiveness?

1.2. The Role of Generative Learning in Answer Justification

Generative learning strategies, which involve actively constructing and interacting with new information, have been shown to enhance learning more effectively than passive study methods like rereading (for a review, see Fiorella & Mayer, 2016). Some of these techniques rely on creating connections between to-be learned material and prior knowledge (Martin & Pressley, 1991; Ozgungor & Guthrie, 2004), oneself (Symons & Johnson, 1997), or mnemonic devices (Putnam, 2015). A subset of generative learning strategies emphasize engagement through elaboration. One such technique, self-explanation (Chi et al., 1989),

requires learners to generate explanations for themselves while learning new material. This process helps learners monitor understanding, identify knowledge gaps, and integrate novel information with prior knowledge, making it more meaningful (Bisra et al., 2018). Research indicates that generating explanations improves future test performance relative to study techniques that do not involve explanation-based elaboration (Tajika et al., 2007; Bisra et al., 2018).

Why is self-explanation effective? One account argues that the benefit of self-explanation on future test performance may be explained by elaborative encoding— by engaging in self-explanation, learners build a robust network of interconnected information, creating multiple retrieval pathways (M. A. McDaniel, 2023). This detailed network increases the likelihood of successful recall. Self-explanation and answer justification are similar in that they require individuals to explain various concepts, but unlike self-explanation, which typically occurs during the acquisition of new information, answer justification takes place at the time of testing. For this reason, its benefit on test performance may be better explained by the elaborative retrieval account. This account suggests that the act of justifying an answer strengthens retrieval by generating additional cues linked to the target memory (M. A. McDaniel, 2023); that is, when justifying their responses, learners may engage in retrieval around additional cues, increasing the chances of recalling the correct answer.

This retrieval-based explanation aligns with the proposed explanation for why MC questions with competitive alternatives—plausible but incorrect responses—lead to increased performance on later tests compared to questions that use noncompetitive alternatives (J. L. Little et al., 2012; J. L. Little & Bjork, 2015). For example, J. L. Little et al. (2019) examined whether participants given well-crafted MC questions spontaneously recall relevant details related to incorrect alternative responses. Participants were given MC trivia questions and prompted to either explain why each alternative response was incorrect (Experiment 1) or justify their chosen answer (Experiment 2). The results showed that participants did spontaneously retrieve information related to incorrect alternatives, and notably those retrievals were associated with increased performance on later cued-recall tests related to those alternatives. If participants are spontaneously retrieving information at test, requiring answer justification may further encourage this process and improve performance both at test and in subsequent tests.

In answer justification, rather than relying on competitive alternatives to encourage deeper thinking, the act of justification itself prompts elaborative retrieval. To understand, consider the MC question: "Which of these would belong in a mineral collection?" with the answer choices *granite*, *basalt*, *quartz*, and *schist*. A student answering traditionally might simply select the most recognizable option, focus only on the first plausible choice, and stop once they feel confident in their selection. However, when required to justify their answer, the student must evaluate each option more thoroughly, potentially retrieving information about minerals, as well as metamorphic and igneous rocks. This elaborate retrieval process increases the likelihood of correctly identifying quartz as the only true mineral in the list and may even strengthen future recall.

Additionally, answer justification may promote effortful retrieval in ways that standard MC questions do not (Greving & Richter, 2022). While elaborative retrieval strengthens memory by increasing the number of retrieval routes to a target memory, MC questions can often be answered through recognition or chance (Greving & Richter, 2022). Unless specifically designed to circumvent this (J. L. Little & Bjork, 2015), learners may not feel motivated to invest significant cognitive effort. In contrast, questions that require justification demand additional processing and provide some structure and accountability, making it

more likely that learners engage in effortful retrieval and generate elaborative information, ultimately enhancing learning outcomes.

1.3. The Role of Metacognition in Answer Justification

Metacognition—the process of evaluating and regulating one's understanding and cognitive performance (Flavell, 1979)—may contribute to improved test performance when MC questions include answer justifications. Effective metacognitive monitoring involves assessing how well information is understood and can directly inform what decisions learners make about their learning process (i.e., metacognitive control, e.g., Dunlosky & Rawson, 2012). Metacognitive monitoring is often measured by having learners make judgments about their learning at various points and comparing these judgments to objective indicators of performance (Nelson & Narens, 1990; Schraw, 2009). For instance, a learner might study material, estimate their likelihood of remembering it on a 0–100 scale, and then take a test. The alignment between their judgment and test performance reflects their metacognitive monitoring accuracy—how well they calibrate confidence to actual performance. These judgments can be made before an assessment (predictions) or after (postdictions) and at varying levels of specificity, such as across an entire test (global judgments) or within specific topics (topic-level judgments).

How might answer justification influence metacognitive accuracy? According to the *cue-utilization framework* (Koriat, 1997), metacognitive judgments are not based on direct access to memory strength but rather on inferences drawn from various cues. By requiring learners to retrieve and articulate their reasoning, answer justification may direct attention to a wider array of meaningful cues about their memory and understanding that are more diagnostic of test performance, rather than superficial cues like familiarity with the learning material (T. D. Griffin et al., 2019a). For example, research has shown that engaging in self-explanation while reading text material improves metacognitive accuracy by helping learners recognize gaps in their knowledge (e.g., T. D. Griffin et al., 2008; Fukaya, 2013). In the context of MC questions, answer justification likely operates similarly—encouraging learners to critically evaluate their responses could lead to a better alignment between perceived and actual comprehension. Furthermore, when people anticipate having to explain their reasoning, they may be more likely to base their predictions on relevant cues, leading to more accurate judgments. For example, research suggests that students tend to align their metacognitive judgments with the type of test they expect, indicating that expectations can shape the way that students evaluate their understanding (T. D. Griffin et al., 2019b; Thiede et al., 2011; Wiley et al., 2008). Thus, MC questions with answer justification may lead to more accurate predictive and postdictive judgments relative to standard MC questions.

1.4. Present Research

Answer justification may offer educators a simple yet effective way to enhance students' recall during tests by encouraging them to self-generate cues that promote elaborative retrieval. Additionally, it may improve metacognitive accuracy, helping students better assess their own understanding. To examine these potential benefits, we analyzed differences in test performance and metacognitive judgments between participants who were required to provide answer justifications for MC questions and those who were not.

Experiment 1 investigated these ideas by having participants study an introductory geology text before being randomly assigned to one of two groups. Both groups provided global and topic-level metacognitive judgments before and after the test, which consisted of 12 MC questions. However, only the experimental group was required to provide answer justifications. We hypothesized that participants who justified their answers would score

higher on the test and that those expecting to justify their answers would demonstrate higher metacognitive accuracy (for both global and topic-level judgments) both before and after testing compared those who answered MC questions without justification.

In Experiment 2, we attempted to replicate the findings of Experiment 1 and examine how answer justification influences subsequent retrieval on a delayed MC test compared to MC testing alone. Similarly, we hypothesized that participants in the answer justification group would outperform participants in the MC-only group on both the immediate and delayed test.

2. Experiment 1

2.1. Methods

2.1.1. Sample

Participants were 110 college students from an open enrollment university in the Intermountain West of the United States. Participants were recruited from a departmental participant pool consisting of undergraduate students who are required to participate in research for course credit. The majority of participants were from Introductory Psychology courses. Participants were excluded if they had inconsistencies in providing metacognitive judgments for either global or topic-level judgments (for example, giving a 0% for a global judgment and then values for topic-level judgments or vice-versa, $n = 13$) or if they admitted to cheating by using outside resources during the test ($n = 18$). After applying these exclusion criteria, the final sample consisted of 79 participants. Participants were randomly assigned to one of two groups: the multiple-choice-only (MC-only) group ($n = 40$) or the answer justification group ($n = 39$). To our knowledge, Experiment 1 was the first experiment of its kind to investigate the influence of answer justification on multiple-choice test performance in a non-classroom setting with geology materials. We attempted to reach approximately 40 participants per group, similar to the group sizes of Papadopoulos et al. (2022) with group sizes of 54 and 44.

2.1.2. Materials

The study was conducted using an online survey administered through Qualtrics. All participants provided informed consent.

2.1.3. Procedure

The full experimental procedure is depicted in Figure 1.

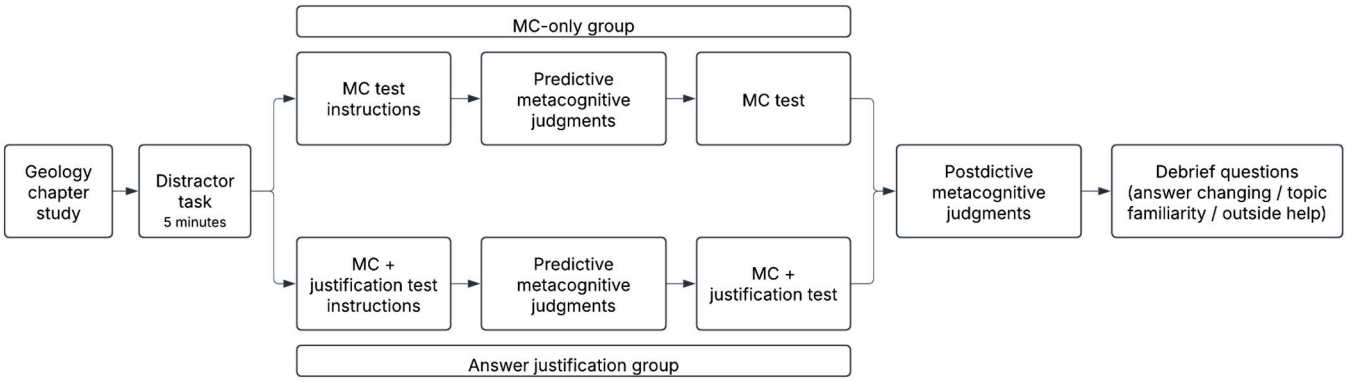


Figure 1. Procedure for Experiment 1.

Learning Phase. Both groups read a chapter from an introductory geology textbook covering fundamental geology concepts (Earle, 2019). The chapter, approximately 3150 words long, included images to illustrate key concepts. Participants were allowed

to read at their own pace. Upon completing the reading, they engaged in a five-minute distractor task, which involved playing an online fishing game.

After the distractor task, participants were informed about the type of test they would take. Those in the MC-only group were told they would complete a standard MC test, whereas those in the answer justification group were instructed that, in addition to answering MC questions, they would need to "explain why they selected" each response.

Predictions. Following these test-specific instructions, participants were asked to estimate their expected test performance. Firstly, they provided an estimate of the percentage of questions (0–100%) they believed they would answer correctly overall (global judgment). After, participants were given four topics for each content subsection of the geology text (*defining geology*, *minerals and rocks*, *plate tectonics*, and *geological time*) and were again asked to estimate the percentage of questions for each topic they would answer correctly (topic-level judgments). Each topic corresponded to a different number of questions, ranging from 2 to 5 questions. The different numbers of questions reflected the varying length of each subsection of the text.

Test. The test consisted of 12 MC questions covering the material from the textbook chapter participants read. Participants in the MC-only group answered each question in a standard multiple-choice format. All questions were presented on a single page, and participants had unlimited time to respond and modify their answers as needed. Participants in the answer justification group received the same question stems and response choices but were also required to provide a brief explanation for each response in a text entry box (Figure 2). The survey was designed to ensure that all participants answered every MC question and, for those in the answer justification group, entered text in the justification box before proceeding. Test questions were presented in a fixed order for all participants.

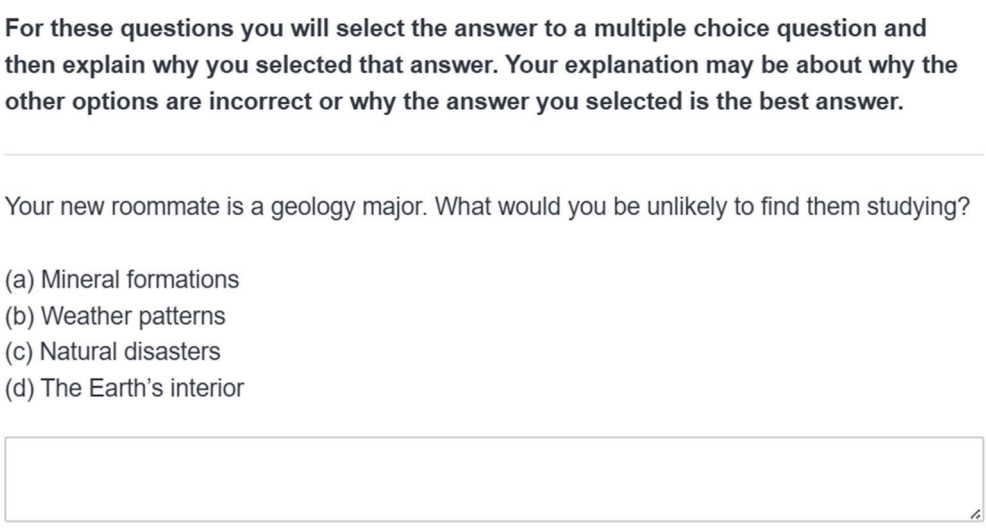


Figure 2. Example of the answer justification test format.

Postdictions and Debrief. After completing the test, participants in both groups provided postdictive global and topic-level judgments, using the same format as their predictions. Before concluding the experiment, participants were asked, "Did you use anything aside from your own memory to answer the questions about geology (for example google, a friend, notes, etc.)?" Participants who answered "yes" to this question were excluded from analyses. Participants were also asked about their familiarity with geology concepts presented in the chapter (on a 5-point scale ranging from "not familiar at all" to "extremely familiar") and if they were in the answer justification group, how frequently they recalled changing their answers (on a 5-point scale ranging from "never" to "always").

2.2. Results

2.2.1. Test Performance

Test performance was measured as the average percentage of correct responses across all 12 MC questions within each group and is presented in Figure 3. For the answer justification group, only the accuracy of the MC responses was considered; written justifications were not evaluated for correctness. The results indicated a significant difference in performance between groups, with the answer justification group scoring higher (M = 64%, SD = 18) than the MC-only group (M = 55%, SD = 16), $t(77) = 2.39$, $p = 0.019$, $d = 0.54$.

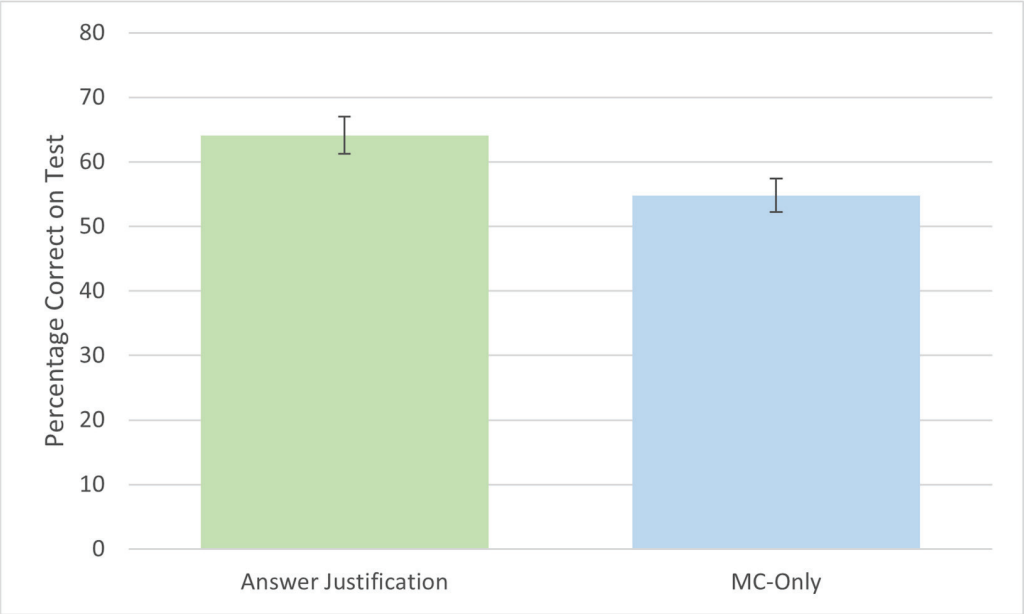


Figure 3. Performance on test for MC-only and answer justification groups in Experiment 1. Error bars depict the standard error of the mean.

2.2.2. Metacognitive Accuracy

Descriptive statistics for all performance and judgment variables are presented in Table 1. We compared judgment accuracy for predictive and postdictive judgments by group. To measure metacognitive accuracy, we computed bias scores (Schraw, 2009). Bias is the difference between one's judgment and actual performance on a given topic, and it can reveal whether participants are overconfident (indicated by positive bias scores) or underconfident (indicated by negative scores) about their performance.

Global Judgments. We calculated bias by subtracting participants' overall performance on the multiple-choice test from their global judgments. A 2 (judgment type: predictive, postdictive) × 2 (group: MC-only, answer justification) mixed ANOVA on bias scores revealed no significant main effect of bias type, $F(1, 77) = 3.12$, $p = 0.081$, $\eta_p^2 = 0.04$, or group, $F(1, 77) = 3.35$, $p = 0.071$, $\eta_p^2 = 0.04$. The interaction between bias type and group was significant, $F(1, 77) = 5.73$, $p = 0.019$, $\eta_p^2 = 0.07$.

Follow-up paired-samples t-tests revealed that this interaction was driven by a significant difference in bias scores between predictive and postdictive bias for the MC-only group, $t(39) = 2.77$, $p = 0.009$, $d = 0.44$, such that predictive bias (M = 9, SD = 21) was higher than postdictive bias ($M = -1$, SD = 24). However, for the answer justification group, there was no significant difference between predictive bias ($M = -5$, SD = 22) and postdictive bias ($M = -3$, SD = 23), $t(38) = 0.48$, $p = 0.635$, $d = 0.08$. Thus, participants in the MC-only group improved their bias scores from predictions to postdictions, whereas participants in the answer justification group had low bias scores at the time of predictions.

Topic-Level Judgments. We computed bias by subtracting participants' actual topic performance from their topic-level judgments. We then computed an average across topics.

A 2 (judgment type: predictive, postdictive) × 2 (group: MC-only, answer justification) mixed ANOVA on topic-level bias scores revealed no significant main effect of judgment type, $F(1, 77) = 2.23$, $p = 0.139$, $\eta_p^2 = 0.03$, or group, $F(1, 77) = 1.49$, $p = 0.188$, $\eta_p^2 = 0.02$. The interaction between judgment type and group was significant, $F(1, 77) = 4.53$, $p = 0.037$, $\eta_p^2 = 0.06$.

Follow-up paired-samples *t*-tests revealed that this interaction was driven by a significant difference between predictive and postdictive bias for the MC-only group, $t(39) = 2.33$, $p = 0.025$, $d = 0.37$, such that predictive bias ($M = 6$, $SD = 24$) was higher than postdictive bias ($M = -2$, $SD = 28$). However, for the answer justification group, there was no significant difference between predictive bias ($M = -5$, $SD = 23$) and postdictive bias ($M = -4$, $SD = 23$), $t(38) = -0.51$, $p = 0.614$, $d = 0.08$; that is, participants in the answer justification group were already fairly accurate (and even slightly underconfident, on average) in their topic-level judgments prior to taking the multiple-choice test.

Table 1. Average magnitude (percentage) of participants' global and topic-level metacognitive judgments and performance in Experiment 1.

		Topic-Level Judgments			
	Global Judgments	**Defining Geology**	**Minerals and Rocks**	**Plate Tectonics**	**Geologic Time**
		MC-Only Group			
Predictive Judgments	64 (3)	77 (4)	62 (4)	56 (4)	45 (4)
Performance	55 (16)	68 (5)	47 (5)	63 (3)	35 (5)
Postdictive Judgments	54 (4)	62 (5)	52 (5)	50 (5)	50 (4)
		Answer Justification Group			
Predictive Judgments	59 (4)	72 (4)	60 (4)	53 (4)	52 (4)
Performance	64 (18)	77 (5)	57 (5)	72 (3)	42 (5)
Postdictive Judgments	61 (4)	71 (5)	60 (5)	55 (4)	56 (4)

Note: Standard errors are in parentheses.

2.2.3. Familiarity Ratings and Self-Reported Answer Changing

Overall, participants reported moderate levels of familiarity with the content. Participants in the answer justification group reported numerically higher ratings of content familiarity ($M = 2.44$, $SD = 0.79$) than the MC-only group ($M = 2.17$, $SD = 0.90$), but this difference was not statistically significant, $t(77) = 1.37$, $p = 0.176$, $d = 0.31$. Participants in the answer justification group were also asked to report how frequently they changed their answers on the test and the majority of participants indicated they changed their answers "sometimes" (see the top of Table 2).

Table 2. How often participants in the answer justification group reported changing their responses to multiple-choice questions as they were writing explanations in Experiments 1 and 2.

	Never	**Sometimes**	**About Half the Time**	**Most of the Time**	**Always**
Experiment 1	28%	54%	10%	8%	0%
Experiment 2	24%	53%	16%	6%	2%

2.3. Discussion

The results of Experiment 1 highlight the effectiveness of answer justification in promoting accurate retrieval during testing. Participants who provided justifications selected

more correct responses on MC questions compared to those who answered standard MC questions without justification.

Additionally, patterns in metacognitive judgment accuracy differed between groups. The MC-only group overestimated their performance before the test but adjusted their postdictions, improving their accuracy. In contrast, the answer justification group showed no significant change in metacognitive accuracy from predictions to postdictions, suggesting that the expectation to justify answers may have led to more realistic performance predictions from the outset. However, our measures of metacognitive accuracy are limited, a point we return to in the General Discussion.

Experiment 1 had several limitations that we aimed to address in Experiment 2. Firstly, we were slightly under-powered to detect the observed effect of answer justification on test performance, so in Experiment 2, we attempted a high-powered replication of this outcome. Secondly, the percentage-based metacognitive judgments may have confused participants and led to some errors in providing judgments. Thus, in Experiment 2, we had participants make their judgments on the same scale as performance to reduce ambiguity in interpreting our bias scores (as recommended by Dunlosky et al., 2016). Further, in Experiment 1, we did not measure time on the test, which could offer insights into the relative efficiency of each test type. This measurement was added in Experiment 2.

3. Experiment 2

The results of Experiment 1, along with previous research, suggest that answer justification improves performance on MC tests compared to standard MC questions. If the elaborative retrieval account is correct and answer justification engages elaborative retrieval processing during testing, it should also improve performance on subsequent tests, functioning as a form of test-potentiated learning or retrieval practice (S. K. Carpenter et al., 2022; Roediger & Butler, 2011). Retrieval practice, a learning strategy that involves actively recalling information from memory, has been shown to enhance long-term retention compared to restudy (Rowland, 2014). More recently, work has begun to explore how retrieval practice, combined with forms of elaborative encoding, may increase learning even more than retrieval practice alone (for a review, see M. A. McDaniel, 2023). While past research demonstrates that MC questions can be effective for retrieval practice (e.g., J. L. Little et al., 2012; Smith & Karpicke, 2014), their benefits depend on question construction. If answer justification encourages elaboration similar to how MC questions with competitive alternatives function (e.g., J. L. Little et al., 2012), then answer justification should also enhance delayed test performance compared to MC-only tests, making it a more effective long-term strategy than standard MC testing.

To investigate this possibility, Experiment 2 sought to replicate the findings of Experiment 1 and assess whether answer justification also functions as an effective encoding strategy, improving retrieval of the same information after a longer delay. Participants underwent the same procedure as in Experiment 1 and then returned for a second experimental session two days later, during which they took a standard multiple-choice test on the same content.

3.1. Methods

3.1.1. Sample

Experiment 2 included 257 undergraduate students from the same university participant pool as Experiment 1. Prior to data collection, a sample size analysis was conducted using G*Power 3.1.9.7 (Faul et al., 2009) for an independent sample *t*-test, assuming 80% power and a medium effect size (d = 0.5, similar to the effect of answer justification on performance observed in Experiment 1), which suggested a sample size of 64 per group,

or 128 participants in total. To ensure adequate power, we continued data collection until 132 participants had completed the second session, at which point we began our analysis.

Participants were excluded if they provided impossible metacognitive judgments (for example, indicating they would answer four questions correctly on a topic with only three corresponding questions) in providing either global or topic-level judgments (n = 19), admitted to cheating (n = 13), or failed to complete the full procedure in Session 1 (n = 2). Due to the multi-day nature of the study, we experienced a significant dropout rate (53%) between the first and second sessions, resulting in a final sample size of 238 participants for Session 1 (n = 126 MC-only, 112 answer justification) and 112 participants for Session 2 (n = 57 MC-only, 55 answer justification).

3.1.2. Materials

The textbook chapter, test questions, and explanations of MC questions and answer justification were identical to Experiment 1. We modified the prediction and postdiction format. Instead of estimating the percentage of correct answers, participants were asked to predict the number of questions they expected to answer correctly (e.g., "*The test you are about to take consists of 12 questions. How many questions do you think you will answer correctly?*"). This adjustment aimed to make participants' self-assessments clearer and more consistent.

3.1.3. Procedure

Experiment 2 consisted of two online sessions. The first session followed the same procedure as Experiment 1, including the learning phase, immediate test phase, and the predictive and postdictive judgments. The second session opened 48 h later, and participants had 12 h to complete the session. During this session, all participants completed a standard MC test. Participants also completed a final (postdictive) judgment at this time, but due to an unfortunate coding error, these data were uninterpretable and are thus not reported here. At the end of the second session, participants were asked the same questions as in Experiment 1 regarding the use of additional resources (e.g., Google, notes, friends) to ensure test integrity. Participants were also timed during the immediate test to examine how long MC-only and answer justification groups took to answer their respective question types (and provide a measure of learning efficiency).

3.2. Results

3.2.1. Test Performance

We first analyzed test performance across both the immediate and delayed tests using a 2 (group: MC-only, answer justification) × 2 (test timing: immediate, delayed) mixed ANOVA. The results showed a main effect of group, $F(1, 110) = 4.70$, $p = 0.032$, $\eta_p^2 = 0.04$, with the answer justification group (M = 67%, SE = 2) outperforming the MC-only group (M = 63%, SE = 2). There was no main effect of test timing, $F(1,110) = 1.86$, $p = 0.175$, $\eta_p^2 = 0.02$, and no interaction, $F(1,110) = 3.50$, $p = 0.064$, $\eta_p^2 = 0.03$. These results suggest that the benefits of answer justification persist over a delay.

We next examined participants' performance in Session 1, which was analyzed using an independent-samples t-test. Consistent with Experiment 1, participants in the answer justification group (M = 67%, SD = 18) outperformed those in the MC-only group (M = 62%, SD = 17), $t(236) = 2.38$, $p = 0.018$, $d = 0.31$. These results successfully replicate the positive effect of answer justification on test performance.

We then analyzed group differences for the 47% of participants who completed the 48 h delayed MC test. Numerically, participants in the answer justification group (M = 67%, SD = 17) scored higher than those in the MC-only group (M = 61%, SD = 17), but the results were only marginally significant $t(110) = 1.82$, $p = 0.071$, $d = 0.34$.

3.2.2. Time on Test

We examined the length of time each group spent on the immediate test. Participants in the answer justification condition took significantly more time to complete the test (M = 14.19 min, SD = 10.52) than participants in the MC-only group (M = 3.98, SD = 3.05), $t(236) = 10.4$, $p < 0.001$, $d = 1.35$. For each participant, we calculated *gains per minute* (GPM; see Rawson & Dunlosky, 2013) by dividing their test performance by the average number of minutes spent during the test. On average, participants in the MC-only group had a GPM of 0.20 (SD = 0.11), while the answer justification group was 0.06 (SD = 0.03). These results indicate that while answer justification increases overall performance, it takes substantially longer for participants to complete.

3.2.3. Metacognitive Accuracy

Descriptive statistics for all performance and judgment variables are presented in Table 3. As a reminder, judgments were made for the corresponding number of questions (e.g., for each topic) rather than on a percentage scale.

Table 3. Average magnitude of participants' global and topic-level metacognitive judgments and performance in Experiment 2.

		Topic-Level Judgments			
	Global Judgments	**Defining Geology**	**Minerals and Rocks**	**Plate Tectonics**	**Geologic Time**
			MC-Only Group		
Predictive Judgments	7.60 (0.19)	1.79 (0.04)	2.02 (0.06)	2.66 (0.10)	1.23 (0.06)
Immediate Test Performance	7.40 (0.18)	1.60 (0.05)	1.61 (0.08)	3.25 (0.10)	0.94 (0.07)
Postdictive Judgments	6.64 (0.20)	1.70 (0.04)	1.69 (0.07)	2.27 (0.11)	1.06 (0.06)
			Answer Justification Group		
Predictive Judgments	7.59 (0.20)	1.83 (0.04)	1.98 (0.07)	2.66 (0.10)	1.29 (0.06)
Immediate Test Performance	8.04 (0.20)	1.70 (0.05)	1.87 (0.08)	3.56 (0.11)	0.92 (0.07)
Postdictive Judgments	7.38 (0.23)	1.80 (0.06)	1.96 (0.08)	2.59 (0.12)	1.10 (0.06)

Note: Standard errors are in parentheses. The immediate test and global judgments were out of 12 questions; topics varied in the number of questions (defining geology: 2 questions, minerals and rocks: 3 questions, plate tectonics: 5 questions, geologic time: 2 questions).

Global Judgments. As in Experiment 1, we computed bias by subtracting participants' performance from their judgments. A 2 (judgment type: predictive, postdictive) × 2 (group: MC-only, answer justification) mixed ANOVA on bias scores revealed a significant main effect of judgment type, $F(1, 236) = 18.12$, $p < 0.001$, $\eta_p^2 = 0.07$, with predictive bias ($M = -0.13$, $SE = 0.17$) being lower than postdictive bias ($M = -0.71$, $SE = 0.16$). The main effect of group was not significant, $F(1, 236) = 0.93$, $p = 0.335$, $\eta_p^2 = 0.004$. The interaction was significant, $F(1, 236) = 7.60$, $p = 0.006$, $\eta_p^2 = 0.03$.

Follow-up paired-samples *t*-tests revealed that this interaction was driven by a significant difference in bias scores between predictive and postdictive bias for the MC-only group, $t(125) = 5.68$, $p < 0.001$, $d = 0.51$, such that predictive bias ($M = 0.21$, $SD = 2.41$) was higher than postdictive bias ($M = -0.75$, $SD = 2.25$). However, for the answer justification group, there was no significant difference between predictive bias ($M = -0.46$, $SD = 2.71$) and postdictive bias ($M = -0.66$, $SD = 2.64$), $t(111) = 0.94$, $p = 0.335$, $d = 0.09$.

Topic-Level Judgments. A 2 (judgment type: predictive, postdictive) × 2 (group: MC-only, answer justification) mixed ANOVA on bias scores revealed a significant main effect of judgment type, $F(1, 236) = 22.37$, $p = < 0.001$, $\eta_p^2 = 0.09$, with predictions showing a very small degree of overconfidence ($M = 0.003$, $SE = 0.04$) and postdictions showing underconfidence ($M = -0.16$, $SE = 0.04$). The main effect of group was not significant, $F(1, 236) = 0.69$, $p = 0.408$, $\eta_p^2 = 0.003$. The interaction between judgment type and group was significant, $F(1, 236) = 5.93$, $p = 0.016$, $\eta_p^2 = 0.03$.

Follow-up paired-samples *t*-tests revealed that this interaction was driven by a significant difference between predictive and postdictive bias for the MC-only group, $t(125) = 5.82$, $p < 0.001$, $d = 0.52$, such that predictive bias ($M = 0.08$, $SD = 0.57$) was higher than postdictive bias ($M = -0.17$, $SD = 0.57$). However, for the answer justification group, there was no significant difference between predictive bias ($M = -0.07$, $SD = 0.70$) and postdictive bias ($M = -0.15$, $SD = 0.70$), $t(111) = 1.43$, $p = 0.164$, $d = 0.14$.

3.2.4. Familiarity Ratings and Self-Reported Answer Changing

Familiarity ratings did not differ for participants in the answer justification group ($M = 2.09$, $SD = 0.89$) or the MC-only group ($M = 2.18$, $SD = 0.90$), $t(110) = 0.61$, $p = 0.542$, $d = 0.12$. Similar to Experiment 1, participants in the answer justification group most frequently stated that they changed their answers "sometimes" (see the bottom of Table 2).

3.3. Discussion

The results of Experiment 2 successfully replicate the findings from Experiment 1, demonstrating that participants in the answer justification group outperformed those in the MC group on the immediate test. Additionally, the results reaffirmed that the predictive and postdictive bias scores varied depending on the testing group. Performance on the standard MC test administered 48 h later showed a numerical advantage for the answer justification group over MC-only, but analyses did not reach significance.

4. General Discussion

Across two experiments, we demonstrated that answer justification is an effective technique for enhancing test performance compared to standard MC questions. In our first experiment, participants who provided justifications were more likely to select the correct response than those who simply answered the questions alone (Figure 3). Our second experiment replicated this finding and explored how answer justification influenced learning after a delay (Figure 4). In the delayed test (48 h later), we found the same numerical effect in the same direction as the immediate test, where participants in the answer justification group scored higher than those who did not have the option to justify their answers, but these results were nonsignificant. Overall, our three comparisons of answer justification to MC-only tests on MC test performance yielded small to medium effect sizes (Experiment 1: $d = 0.54$; Experiment 2, immediate yest: $d = 0.31$, delayed test: $d = 0.34$). These results suggest that answer justification may involve elaborative retrieval, but further exploration is needed.

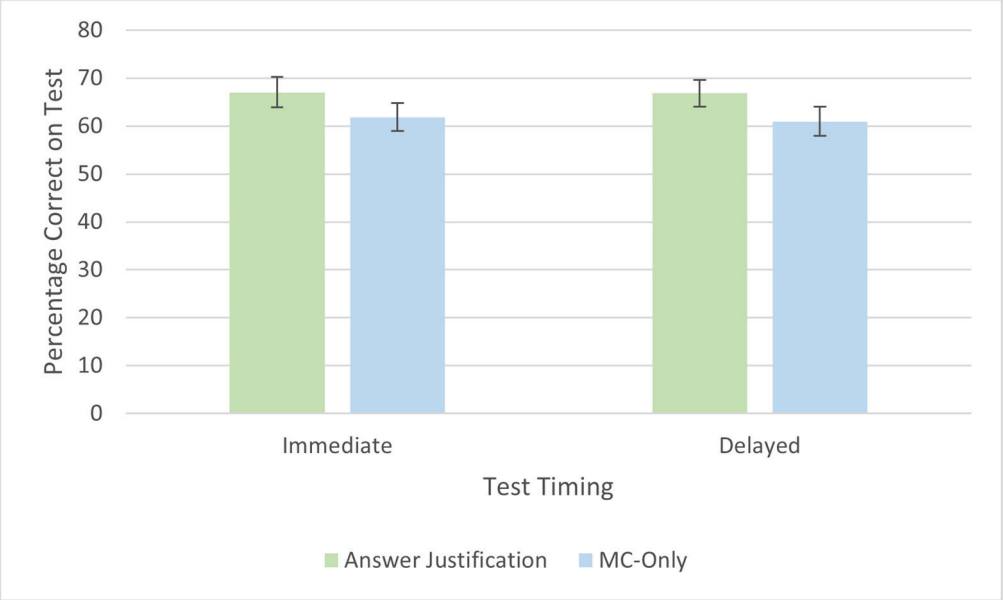


Figure 4. Performance on immediate test and delayed test in Experiment 2. Error bars depict the standard error of the mean.

4.1. Metacognitive Implications

Our findings provide some initial evidence that anticipating an answer justification task may influence metacognitive judgments (see Tables 1 and 3). Participants in the MC-only group exhibited a significant shift in their predictive and postdictive accuracy (in both global and topic-level judgments), initially overestimating their performance and becoming more accurate (or even underconfident) in their postdictions. This outcome is consistent with previous research showing a *postdiction superiority effect*, likely driven by retrieval-based processes that occur during a test (e.g., remembering what happened when questions were answered; Pierce & Smith, 2001; Rivers et al., 2019). In contrast, participants in the answer justification group showed more stable metacognitive judgments. Interestingly, participants who anticipated and completed answer justifications tended to exhibit underconfidence in both global and topic-level judgments.

However, because our measure of metacognitive accuracy (bias) is a derived measure that includes both judgment magnitude and test performance, our metacognitive outcomes should be interpreted with caution; that is, we cannot conclude that differences in bias scores are due to better monitoring ability for the answer justification group (Dunlosky et al., 2016). Thus, further systematic investigation will be needed to evaluate the influence of answer justification on metacognitive monitoring ability. Specifically, future research should have learners make metacognitive judgments for each multiple-choice question (with or without answer justification), which would allow for a cleaner measure of bias and a calculation of relative metacognitive accuracy (i.e., how well learners can discriminate between material they know well versus not as well).

4.2. Alternative Explanations

While our findings provide some support for a retrieval-based explanation, alternative explanations for the benefits of answer justification should be considered. Some researchers suggest that answer justification may reduce test anxiety, as students often prefer question formats that allow them to explain their reasoning rather than simply selecting an answer (Nield & Wintre, 1986). One potential source of anxiety in MC questions is a mismatch between a student's knowledge and the given answer choices. Providing a justification option may give students a greater sense of control, reducing test-related stress (McKeachie

et al., 1955). In support of this idea, students have expressed preference for question formats that allow them to more fully articulate what they have learned—such as essay and short-response—even while acknowledging that MC questions are "easier" (Nield & Wintre, 1986). However, test anxiety alone is unlikely to explain our results. Unlike prior studies conducted in high-stakes classroom settings, our research took place in a low-stakes laboratory environment, where anxiety was presumably lower. Moreover, it is possible that answer justification increased effort rather than reduced stress.

Another possible explanation is that better performance on the delayed test in Experiment 2 may be attributed to retrieval success rather than elaborative retrieval. Previous work on the retrieval practice has demonstrated that success during retrieval practice strongly predicts later test performance (Pan & Rickard, 2018; Rowland, 2014, though see Chan et al., 2024 for another perspective). In our study, participants in the answer justification group for Experiment 2 may have performed better on the delayed test simply because they answered more questions correctly on the immediate test, rather than due to more elaborative retrieval related to items brought to mind during justification. Examining the results of Experiment 2, participants in both the answer justification group and MC-only group did show a strong, positive correlation between performance at immediate and delayed tests ($r = 0.80$, $p < 0.001$, $r = 0.77$, $p < 0.001$, respectively). These correlations suggest that retrieval success had some relationship with later performance, though such a relationship was similar across both groups and may not explain the entire benefit of answer justification.

4.3. Limitations and Future Directions

Several methodological limitations should be acknowledged. Firstly, data collection was conducted entirely online via Qualtrics, which limited our ability to control or monitor participant behavior. Although participants were asked to self-report cheating, it is possible that some undisclosed cheating occurred, affecting the outcomes. Secondly, there was a substantial dropout rate in Experiment 2 (53% of participants did not complete the delayed test), which reduced our sample size for the second session and limited the generalizability of our results. Further, our metacognitive questions changed from Experiment 1 to Experiment 2. We made this change to address errors in participant responding, but this change in measurement across the two studies makes them challenging to compare.

Future work could investigate other applications of answer justification as well as its limitations. One avenue of research could explore how using answer justification as an encoding technique influences performance on delayed *transfer* tests. Much like J. L. Little et al. (2019), if engaging in justification causes participants to spontaneously retrieve details surrounding the question or related responses, that elaborative retrieval could lead to enhanced performance on related transfer questions at a later test.

Another important direction for future research is examining individual differences to determine how they influence the effectiveness of answer justification. In our experiments, participants reported moderate levels of familiarity with the geology content. However, self-reported familiarity is limited, and an important direction for future research would be to explore the interaction between answer justification and prior knowledge (as it plays a role in other learning techniques, e.g., Witherby & Carpenter, 2022). Understanding how different learner characteristics affect the benefits of justification could help tailor instructional strategies to diverse student populations. Additionally, research should compare answer justification to other generative learning strategies, such as self-explanation (Renkl & Eitel, 2019), elaborative interrogation (e.g., M. A. McDaniel & Donnelly, 1996), and retrieval practice with open-ended responses, to better understand its relative effectiveness and underlying mechanisms. Further comparisons could also be made with social learning

strategies like *peer instruction*, in which students first answer multiple-choice questions individually and then discuss their responses with peers before re-answering (Crouch & Mazur, 2001; Tullis & Goldstone, 2020). This comparison could help determine whether collaborative reasoning provides similar or additional benefits compared to individual justification.

Finally, future studies should examine classroom implementation factors to better understand how answer justification functions in real-world educational settings, including high-stakes exams, online learning platforms, and adaptive testing environments. Understanding how factors like time constraints, instructional support, and grading considerations influence its effectiveness could help educators integrate this technique more effectively. By addressing these questions, future research can refine our understanding of when and how answer justification is most beneficial for learning and assessment.

5. Conclusions

Answer justification is a simple addition to MC questions that can be easily adopted by instructors. Our results indicate that adding answer justification enhances performance during testing. Unlike other best practices for MC tests, which emphasize time spent developing questions with appropriate competitive responses, answer justification may be adopted for a wide variety of question types to introduce an added element of difficulty and retrieval.

While answer justification offers clear benefits, its adoption in classroom settings requires careful planning. One consideration is the increased time required for completion, as answer justification takes longer than standard multiple-choice questions. In Experiment 2, we found that participants spent significantly more time on answer justification questions than MC-only questions (about 10 more minutes on average). Instructors may need to adjust test length or allocate additional time for justifications. Another factor is question suitability; answer justification probably works best for conceptual and application-based questions rather than simple definition-based recall. For example, a question like "What is a mineral?" may not be as effective in this format as "Which of the following is a mineral?" because the latter allows for a more meaningful explanation.

Finally, in classroom settings, answer justifications provide further insight into student thinking compared to standard MC. Students may spontaneously reveal misconceptions, gaps in their thinking, or admit lack of knowledge in their responses. These rich responses provide an opportunity for individualized interventions to address misunderstandings.

Author Contributions: Conceptualization, S.A.C., M.L.R., and A.L.O.; methodology, S.A.C., M.L.R., and A.L.O.; software, S.A.C. and A.L.O.; formal analysis, S.A.C., M.L.R., and A.L.O.; investigation, S.A.C. and A.L.O.; writing—original draft preparation, S.A.C., M.L.R., and A.L.O.; writing—review and editing, S.A.C., M.L.R., and A.L.O.; visualization, S.A.C. and A.L.O.; supervision, A.L.O.; project administration, A.L.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Utah Valley University (protocol code 1215 and approved on 11 September 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are openly available in the Open Science Framework (OSF) at https://osf.io/x9hmv/?view_only=e2b80cfff261469a9b531146f815ef37 (accessed on 10 February 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MOOC	Massive open online courses
GPM	Gains per minute
MC	Multiple-choice

References

Bassett, M. H. (2016). Teaching critical thinking without (much) writing: Multiple-choice and metacognition. *Teaching Theology & Religion*, *19*(1), 20–40. [CrossRef]

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, *30*(3), 703–725. [CrossRef]

Brown, G. T., & Abdulnabi, H. H. (2017). Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Frontiers in Education*, *2*, 24. [CrossRef]

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, *1*(9), 496–511. [CrossRef]

Chan, J. C. K., Davis, S. D., Yurtsever, A., & Myers, S. J. (2024). The magnitude of the testing effect is independent of retrieval practice performance. *Journal of Experimental Psychology: General*, *153*(7), 1816–1837. [CrossRef]

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. [CrossRef]

Crouch, C. H., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, *69*(9), 970–977. [CrossRef]

Dodd, D. K., & Leal, L. (1988). Answer justification: Removing the "trick" from multiple-choice questions. *Teaching of Psychology*, *15*(1), 37–38. [CrossRef]

Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory: Problems and pitfalls. In J. Dunlosky, & S. K. Tauber (Eds.), *The Oxford handbook of Metamemory* (pp. 23–38). Oxford University Press.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271–280. [CrossRef]

Earle, S. (2019). *Physical geology* (2nd ed.). BCcampus. Available online: https://opentextbc.ca/physicalgeology2ed (accessed on 1 March 2023).

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. [CrossRef] [PubMed]

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*(4), 717–741. [CrossRef]

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911. [CrossRef]

Fukaya, T. (2013). Explanation generation, not explanation expectancy, improves metacomprehension accuracy. *Metacognition and Learning*, *8*(1), 1–18. [CrossRef]

Gorrell, J., & Cramond, B. (1988). Students' attitudes toward and use of written justifications for multiple-choice answers. *Educational and Psychological Measurement*, *48*(4), 935–943. [CrossRef]

Greving, S., & Richter, T. (2022). Practicing retrieval in university teaching: Short-answer questions are beneficial, whereas multiple-choice questions are not. *Journal of Cognitive Psychology*, *34*(5), 657–674. [CrossRef]

Griffin, T. D., Mielicki, M. K., & Wiley, J. (2019a). Improving students' metacomprehension accuracy. In J. Dunlosky, & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 619–646). Cambridge University Press. [CrossRef]

Griffin, T. D., Wiley, J., & Thiede, K. W. (2019b). The effects of comprehension-test expectancies on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(6), 1066–1092. [CrossRef]

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, *36*, 93–103. [CrossRef]

Koretsky, M. D., Brooks, J. D., & Higgins, A. D. (2016). Written justifications to multiple-choice concept questions during active learning in class. *International Journal of Science Education*, *38*(11), 1747–1765. [CrossRef]

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. [CrossRef]

Kottke, J. L. (2001). Students' reactions to written test item rebuttals. *Journal of Instructional Psychology*, *28*(4), 256–261.

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, *43*(1), 14–26. [CrossRef]

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*(11), 1337–1344. [CrossRef] [PubMed]

Little, J. L., Frickey, E. A., & Fung, A. K. (2019). The role of retrieval in answering multiple-choice questions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(8), 1473–1485. [CrossRef]

Martin, V. L., & Pressley, M. (1991). Elaborative-interrogation effects depend on the nature of the question. *Journal of Educational Psychology*, *83*(1), 113–119. [CrossRef]

McDaniel, M. A. (2023). Combining retrieval practice with elaborative encoding: Complementary or redundant? *Educational Psychology Review*, *35*(3), 75. [CrossRef]

McDaniel, M. A., & Donnelly, C. M. (1996). Learning with analogy and elaborative interrogation. *Journal of Educational Psychology*, *88*, 508–519. [CrossRef]

McKeachie, W. J., Pollie, D., & Speisman, J. (1955). Relieving anxiety in classroom examinations. *The Journal of Abnormal and Social Psychology*, *50*(1), 93–98. [CrossRef]

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation* (Vol. 26, pp. 125–173). Academic Press. [CrossRef]

Nield, A. F., & Wintre, M. G. (1986). Multiple choice questions with an option to comment. *Teaching of Psychology*, *13*(4), 196–199. [CrossRef]

Overono, A. L. (in press). Exploring the impact of required justifications in multiple-choice elaboration questions on student experiences and performance. *Journal of Scholarship of Teaching and Learning*.

Ozgungor, S., & Guthrie, J. T. (2004). Interactions among elaborative interrogation, knowledge, and interest in the process of constructing knowledge from text. *Journal of Educational Psychology*, *96*(3), 437–443.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *67*(3), 215–227. [CrossRef] [PubMed]

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. [CrossRef] [PubMed]

Papadopoulos, P. M., Obwegeser, N., & Weinberger, A. (2022). Let me explain! The effects of writing and reading short justifications on students' performance. *Journal of Computer Assisted Learning*, *38*(2), 327–337. [CrossRef]

Pierce, B. H., & Smith, S. M. (2001). The postdiction superiority effect in metacomprehension of text. *Memory & Cognition*, *29*, 62–67. [CrossRef]

Putnam, A. L. (2015). Mnemonics in education: Current research and applications. *Translational Issues in Psychological Science*, *1*(2), 130–139. [CrossRef]

Rawson, K. A., & Dunlosky, J. (2013). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology: General*, *142*(4), 1113–1129. [CrossRef]

Renkl, A., & Eitel, A. (2019). Self-explaining: Learning about principles and their application. In J. Dunlosky, & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 528–549). Cambridge University Press. [CrossRef]

Rivers, M. L., Dunlosky, J., & Joynes, R. (2019). The contribution of classroom exams to formative evaluation of concept-level knowledge. *Contemporary Educational Psychology*, *59*, 101806. [CrossRef]

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. [CrossRef]

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. [CrossRef]

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*, 33–45. [CrossRef]

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, *22*(7), 784–802. [CrossRef] [PubMed]

Sparck, E. M., Bjork, E. L., & Bjork, R. A. (2016). On the learning benefits of confidence-weighted testing. *Cognitive Research: Principles and Implications*, *1*, 1–10. [CrossRef] [PubMed]

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, *121*(3), 371. [CrossRef] [PubMed]

Tajika, H., Nakatsu, N., Nozaki, H., Neumann, E., & Maruno, S. (2007). Effects of self-explanation as a metacognitive strategy for solving mathematical word problems. *Japanese Psychological Research*, *49*(3), 222–233. [CrossRef]

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology*, *81*(2), 264–273. [CrossRef]

Tullis, J. G., & Goldstone, R. L. (2020). Why does peer instruction benefit student learning? *Cognitive Research: Principles and Implications*, *5*, 1–12. [CrossRef]

Wiley, J., Griffin, T. D., & Thiede, K. W. (2008). To understand your understanding, one must understand what understanding means. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*. Available online: https://escholarship.org/uc/item/4xb0r6jw (accessed on 10 February 2025).

Witherby, A. E., & Carpenter, S. K. (2022). The rich-get-richer effect: Prior knowledge predicts new learning of domain-relevant information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(4), 483–498. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Does Covert Retrieval Benefit Adolescents' Learning in 8th Grade Science Classes?

Amber E. Witherby [1], Paige E. Northern [2] and Sarah K. Tauber [3,*]

[1] Department of Psychological Sciences, Creighton University, Omaha, NE 68178, USA; amberwitherby@creighton.edu
[2] Department of Psychology and Counseling, Southeast Missouri State University, Cape Girardeau, MO 63701, USA; penorthern@semo.edu
[3] Department of Psychology, Texas Christian University, Fort Worth, TX 76129, USA
* Correspondence: uma.tauber@tcu.edu

Abstract

Retrieval practice can benefit students' long-term learning. However, no prior investigations have explored the degree to which response mode—overt versus covert retrieval—moderates the impact of retrieval practice on adolescents' learning in a classroom context. To explore this issue, 8th grade students learned science concepts that were required for their class. They learned terms in their middle school classrooms by recalling and writing them down (overt retrieval), mentally recalling (covert retrieval), or restudying definitions. They practiced each strategy in 5 sessions and took final tests 2 days later. The impact of covert retrieval on students' learning was similar to that of restudy, and both covert retrieval and restudy were less beneficial relative to overt retrieval. Treatment package effectiveness differed somewhat between students and terms. These outcomes are generally consistent with the retrieval dynamics hypothesis and reveal that response mode can impact the effectiveness of retrieval practice in middle school classrooms.

Keywords: covert retrieval; overt retrieval; adolescents; STEM education

1. Introduction

Retrieval practice involves bringing concepts to mind from memory, and it can support students' long-term retention of information (Roediger & Karpicke, 2006a; for reviews, see Adesope et al., 2017; Carpenter et al., 2022; McDermott, 2021; Rowland, 2014). For instance, students have shown enhanced learning of STEM concepts following retrieval practice (e.g., Darabi et al., 2007; Kang et al., 2011; Pan et al., 2016, 2019), and retrieval practice can benefit learning in educational contexts (McDaniel et al., 2007; for reviews, see Agarwal et al., 2021; Moreira et al., 2019). Most of these investigations focused on learning in higher education. Even so, retrieval practice can also enhance adolescents' learning (for reviews, see Agarwal et al., 2012; Roediger & Karpicke, 2006b). However, *how* students engage in retrieval may be critical, and no research has evaluated if covert (mental) retrieval is effective for adolescents' long-term learning. Our goal was to investigate this issue by recruiting 8th grade science students, many of whom were from underrepresented groups.

Retrieval practice can be beneficial for middle school students' learning (Coyne et al., 2015; Fritz et al., 2007; McDaniel et al., 2011, 2013; McDermott et al., 2014; Roediger et al., 2011). For instance, McDaniel et al. (2011) had 8th grade science students take low-stakes quizzes to encourage retrieval practice for facts from the class content. Students' exam performance was better for facts that appeared on the quizzes relative to facts that did not.

Adolescents also self-report that retrieval practice reduces test anxiety and helps them learn (Agarwal et al., 2014). However, students can engage in retrieval practice in a variety of ways, and the mode of retrieval may moderate the effectiveness of this study strategy.

Most retrieval practice research requires that students generate a response by typing, writing, or speaking their answers. This overt retrieval approach is logical because researchers can evaluate the quality of students' responses during learning, and it simulates some practices used by students. For instance, during group study, students can quiz each other on class concepts and respond aloud. College students also report responding to questions in class when their teachers use iClickers (Witherby & Tauber, 2019). However, students do not always generate an overt response when recalling concepts during study or answering questions during class. Instead, they recall information from memory without producing an explicit response, which is referred to as covert retrieval. When asked how students decide if they have correctly answered a question during study, few college students report having a partner check their responses or writing down answers correctly, both of which require an overt response (Wissman et al., 2011). As well, when a teacher asks a question of the class, often one or two students provide an overt response while the rest of the class mentally recalls their answers. As a final example, when studying by oneself in a quiet environment such as a library, students likely mentally retrieve answers to avoid disrupting others. Consistent with these examples, college students self-report being highly likely to use both covert and overt retrieval when studying (Witherby et al., 2025). Thus, it is important to evaluate the degree to which covert retrieval supports students' long-term learning in educational contexts.

Mixed outcomes have been found for the impact of covert retrieval on college-aged students' learning. Laboratory research has revealed response mode (overt versus covert) during retrieval to have little impact on learning (Higham et al., 2023; Putnam & Roediger, 2013; Smith et al., 2013). Smith et al. (2013) had students learn lists of words by category (e.g., types of vegetables) via overt retrieval, covert retrieval, no retrieval (or restudy), or restudy. A mini meta-analysis including their outcomes and others with similar methodology (Putnam & Roediger, 2013) showed retrieval practice to have a substantial impact on students' learning (d = 1.10) when compared with restudy or no retrieval conditions. By contrast, mode of retrieval—overt versus covert—did not differentially impact students' learning (d = −0.0027). Expanding on these outcomes, Sundqvist et al. (2017) reported a meta-analysis that included outcomes from Jönsson et al. (2014). They also found a low effect size (d = 0.07) comparing overt and covert retrieval, and thus concluded that overt and covert retrieval similarly impact learning.

Other evidence suggests that covert retrieval is inferior relative to overt retrieval for college-aged students' learning (Jönsson et al., 2014; Abel & Roediger, 2018; Kubik et al., 2020; Tauber et al., 2018). Tauber et al. (2018) tasked college students with learning definitions for key terms. Students studied the key term definitions by practicing covert retrieval, overt retrieval, or restudying the information. Meta-analyses showed that covert retrieval was less beneficial for performance on a delayed test relative to overt retrieval, and covert retrieval did not differ from restudy. In a related literature, metacognitive researchers have found analogous outcomes (Ariel et al., 2021; Kubik et al., 2022; Tauber et al., 2015). In this work, overt retrieval practice is contrasted with restudy and conditions for which covert retrieval is presumed to occur. Under covert retrieval conditions, participants make judgments of learning (JOLs) after a delay and with only a cue available (for a review, see Rhodes & Tauber, 2011). JOL conditions are typically less beneficial for retention on delayed tests relative to overt retrieval (but see Rhodes & Tauber, 2011).

Differences between covert and overt retrieval may be particularly pronounced in applied learning contexts. In the only study that has evaluated this possibility, Sumeracki

and Castillo (2022) had college students learn information in a classroom context that was required for their course (Experiment 2) or not (Experiment 1). Students learned the material by using overt retrieval, covert retrieval, restudy, or no retrieval (i.e., control condition without additional exposure to the information). As expected, performance on delayed tests was higher following overt retrieval relative to restudy or no retrieval. By contrast, covert retrieval did not differ from restudy; however, performance did not differ between covert and overt retrieval practice. Extending these outcomes to adolescents in middle school is challenging. Middle school classrooms substantially differ from college classes (e.g., classroom management requirements), and given that middle school students presumably engage in covert retrieval during class and when studying on their own, it is important to evaluate their ability to use covert retrieval to support their learning.

Our research is guided by theory in student learning—specifically, the retrieval dynamics hypothesis (Tauber et al., 2018). From the retrieval dynamics hypothesis, the demand that students face during mental retrieval impacts the degree to which retrieval practice will enhance their learning. From this perspective, students are most likely to practice thorough retrieval for simple information or short answers. In these instances, response modality (overt vs. covert) may have no impact or only a minor impact on enhanced learning from retrieval practice. By contrast, when mentally retrieving complex concepts, students can end retrieval without bringing to mind all parts of the concepts, thereby reducing the impact of retrieval practice on learning. The retrieval dynamics hypothesis suggests that this is most likely to occur during covert retrieval practice due in part to poor indications of how much students have successfully retrieved during retrieval practice. We evaluated this possibility with adolescents by using complicated concepts in the experiment.

To preview the novel reported experiment, we investigated the degree to which covert retrieval enhances adolescents' long-term learning. To explore adolescents' learning in middle school classrooms, we had students learn terms and definitions for their science class. As important, we included three treatment packages that offer practice with covert retrieval, overt retrieval, or restudy to add to the growing literature on methods to enhance adolescents' STEM education. The treatment packages required that students study the definitions of some terms and practice retrieval (overt or covert) for others. Students took short-answer and multiple-choice tests. Consistent with prior research, we expected the overt retrieval treatment package to lead to superior test performance relative to the restudy treatment package (Adesope et al., 2017; Carpenter et al., 2022; McDermott, 2021; Rowland, 2014). It was an open question as to the impact of the covert retrieval treatment package on adolescents' learning in a classroom context; however, from the retrieval dynamics hypothesis and based on recent research (Jönsson et al., 2014; Abel & Roediger, 2018; Kubik et al., 2020; Tauber et al., 2018), we expected the covert retrieval treatment package to be less effective relative to the overt retrieval treatment package.

2. Materials and Methods

2.1. Educational Context

We conducted the experiment in classrooms at a public middle school. The middle school has a population of approximately 400 students, includes grades 6–8, and primarily serves historically underserved youth. According to a 2020–2021 Texas Education Agency School Report Card, 93.9% of the student body was economically disadvantaged. A total of 77.3% of the student population identified as Hispanic, 17.1% as African American, 4.2% as White, 0.5% as Asian, 0.7% as multiethnic, and 0.2% as American Indian. According to the 2021 Federal Report Card for the school, only 28% of 8th grade students performed at grade level or above in science by the end of the course on the State of Texas Assessments of Academic Readiness (STAAR; 49% approached grade level, 7% mastered grade level).

This report identified the school as high poverty with multiple teacher challenges: 18.2% were inexperienced, 12.6% were teaching with emergency or provisional credentials, and 32.9% were teaching subjects for which they were not certified.

To conduct this experiment, we worked with one science teacher and her classes. The teacher identified as 33 years old, White, and as a woman. She was in her first year of teaching general science courses to 8th grade students at the middle school. She held a doctoral degree in Science Education and reported being 41–60% confident (on a categorical scale using 0–20%, 21–40%, 41–60%, 61–80%, and 81–100%) in her knowledge of strategies that are effective for student learning. The teacher and the principal of the school provided consent for the experiment to be conducted in students' classrooms. As well, guardians provided consent (documentation provided in English or Spanish), and students assented to take part in the experiment. The experiment was approved both by the Texas Christian University IRB and the Fort Worth Independent School District. All students were provided with a $20 gift card to compensate them for their participation, as was the teacher ($100 gift card).

2.2. Design and Participants

Treatment package (i.e., overt retrieval, covert retrieval, and restudy) was manipulated using a within-participant design so that all students in all classes practiced all strategies. Ninety-four 8th grade students enrolled in one of four general science classes were invited to participate in the experiment. The sample size was based on the number of students enrolled in the available classes with the goal of obtaining an adequate number of observations to detect the within-participant effect of treatment package on final performance anticipating potential missing data across sessions. A group size of 21 was estimated using G*Power 3 (Faul et al., 2007). Of the 94 students, 24 did not return the guardian consent form, assent form, or both, and 5 students failed to complete the final tests. Thus, the final sample size for all analyses was 65 students (n = 12 in the 9 a.m. class, n = 15 in the 12 p.m. class, n = 18 in the 1 p.m. class, and n = 20 in the 2 p.m. class). This obtained sample size exceeded the sample size estimate (N = 28) from a power analysis that would be necessary to detect the medium-sized within-participant effect of treatment package (f = 0.25, power at 0.80, alpha of 0.05) that was consistent with our predictions (Faul et al., 2007).

Students completed an extensive demographic questionnaire. All students reported being in 8th grade and planning to attend high school. Most students (n = 55, 83.3%) reported planning to attend college (n = 1 only attending if accepted, n = 7 not planning on attending college, n = 3 unsure). One student reported a diagnosis of Autism Spectrum Disorder (n = 4 did not respond), and 8 reported a learning disability diagnosis (n = 6 did not respond). The reported disabilities were dyslexia (n = 4), attention deficit hyperactivity disorder (n = 2), both dyslexia and attention deficit hyperactivity disorder (n = 1), and obsessive compulsive disorder (n = 1). Many students (n = 31, 46.9%) were expecting to get a B in their general science class (n = 15 expecting an A, n = 5 expecting an A or B, n = 1 expecting an A, B, or C, n = 1 expecting an A or C, n = 1 expecting an B or C, n = 3 expecting a C, n = 8 did not respond). In terms of self-reported GPA, several students (n = 28, 42.4%) reported that they earn an equal number of As and Bs (n = 2 mostly As, n = 4 mostly Bs, n = 17 an equal number of Bs and Cs, n = 4 mostly Cs, n = 11 reported another grade combination).

Students reported that the primary language(s) spoken at home was English (n = 31, 46.9%), Spanish (n = 23, 34.8%), or both English and Spanish (n = 7, n = 5 who did not respond). Some students reported living with both parents (n = 25, 37.9%) or multiple adults (combinations of parent(s), aunt/uncle, cousin, roommate, niece, sister-in-law, and grandparent(s), n = 24, 36.4%). A total of 10 students reported living in a single-adult

(mother or grandmother) household, and fewer (n = 6) reported living with a parent and step-parent (n = 1 who did not respond). Few students reported that their mother completed high school (n = 20, 30.3%) or earned a GED (n = 1, n = 15 did not know, n = 5 did not respond), and 6 of these students reported that their mother completed college (n = 6 did not know). Similarly, few students reported that their father completed high school (n = 15, 22.7%, n = 20 did not know, n = 6 did not respond), and 7 of these students reported that their father completed college (n = 2 did not know).

On average, the students were 13.96 years old (SD = 1.11). Forty-seven percent identified as a girl and 51.5% as a boy (1 participant did not respond). Regarding ethnicity, 68.2% identified as Hispanic, 18.2% as Black, 9.1% as Multiethnic, 3% as Asian or Pacific Islander, and 1.5% as White. These demographic characteristics did not differ between the four science classes ($F < 1$, $p = 0.826$, χ^2s ≤ 6.72, ps ≥ 0.348) (see online supplement for demographic data for each class—https://osf.io/p4cq8, accessed on 16 June 2025).

2.3. Materials

We collaborated with the teacher to create all experimental materials. We targeted concepts that students needed to master for their science class and for their state exams. We selected 15 earth science terms (e.g., *jet stream*) and definitions (*a narrow band of strong, fast winds high in the upper troposphere*) from the students' middle school science textbook that were important for understanding the functioning of Earth's oceans and atmospheres, as well as for illustrating the principles of energy and matter and of cause-and-effect (e.g., *meteorology, convection*; all materials provided in the online supplement). The research team drafted a definition of each concept, and the teacher evaluated them for correctness, meeting the goals of course, and test relevancy. As well, the teacher verified that the information had not yet been taught and was presented using language that was suitable for the students' comprehension capabilities. All materials are available in the online supplement (https://osf.io/p4cq8, accessed on 16 June 2025).

2.4. Procedure

The experiment consisted of 6 sessions (see Figure 1), each on a different day during students' regularly scheduled class times at their middle school. Five sessions were on consecutive school days (i.e., Monday–Friday), and the final 6th session was on the following Monday (i.e., after a 2-day delay). Students participated in each session as a group with their classmates. Thus, students in the same class saw the same items in the same order and performed the same learning tasks per session. Even so, tasks were completed individually and without input from peers. During each session, 2 to 5 researchers were present who circulated around the room to ensure that students followed instructions, paid attention, and questions were answered.

For all sessions, instructions were provided on PowerPoint slides, they were read aloud, and time was provided to ask questions. Similarly, all to-be-learned materials were presented on PowerPoint slides and were read aloud by a researcher. In session 1, students were provided with a bird's eye view of the activities. Students were then informed that they would study 15 terms and definitions, and they would practice learning them in three different ways. They were told that they would take a practice test and write the definition for some terms, take a practice test and silently think of the answer for some terms, and restudy the definitions for other terms. Students then received the following instructions:

> "We selected **15 science terms** that you need to learn for your science class. Specifically, you will learn terms that are important for understanding Earth's oceans and atmosphere. They are also important for learning about energy and matter as well as cause and effect.

Your task is to **learn the definition** to each term so that you will be able to remember it on a future memory test. The test will be short-answer format, which means that you will be given the terms and will be asked to write the definition to each.

It is very important that you try your hardest to learn each definition. Remember, these are concepts that you need to learn for your class, so if you work hard at learning them now, you may benefit later! You will be given 15 s to study each definition, which we will present on the screen and read aloud. Do you have any questions before we begin?"

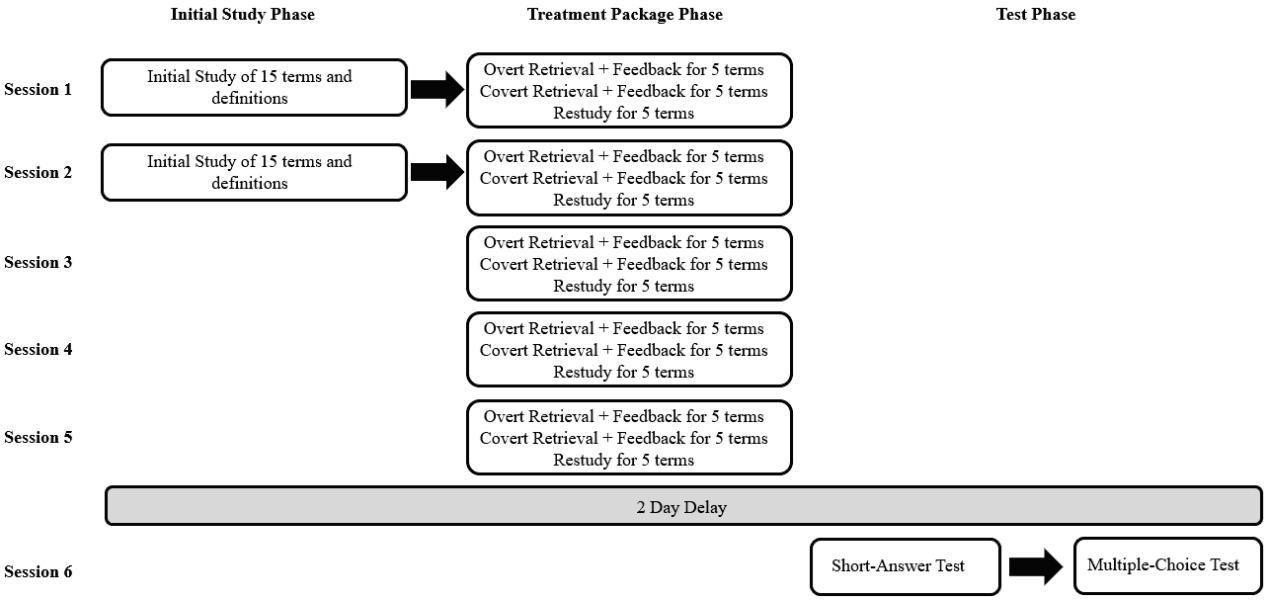


Figure 1. Overview of all experimental sessions. The procedure for session 2 was identical to that of session 1. The initial study phase refers to the first time terms and definitions were studied in that session, which was eliminated after session 2. The treatment package phase was maintained for sessions 3–5, and the procedure for sessions 3–5 was identical. Following a 2-day delay, during session 6, all students completed the final tests and demographic form.

Next, using the time feature for slide transitions in PowerPoint, students studied each term and definition one at a time for 15 s (see Figure 2, left panel). Terms and definitions were presented in a random order for each class. After students studied the 15 terms and definitions, they began the treatment package phase (see Figure 2, right panel). Terms were randomly assigned to treatment package for session 1, and students continued to practice the terms using the same treatment package for all subsequent sessions. For instance, if the term *jet stream* was assigned to the overt retrieval treatment package, then all students in that class would practice overt retrieval for this concept for all sessions. We used the four classes to counterbalance the order of treatment packages.

Treatment package was blocked during learning and the order was counterbalanced across classes. We used this strategy to prioritize altering the order of treatment packages across classes and sessions to ensure that students were exposed to the treatment packages in different orders. This was critical due to the within-participant manipulation of treatment package. Put differently, our strategy ensured that conclusions drawn from the final test data were not dependent upon a specific order of treatment package. It should be noted, however, that the treatment order was imperfect given that data were collected from four classes, across 5 sessions, and 6 treatment package orders (see Table 1).

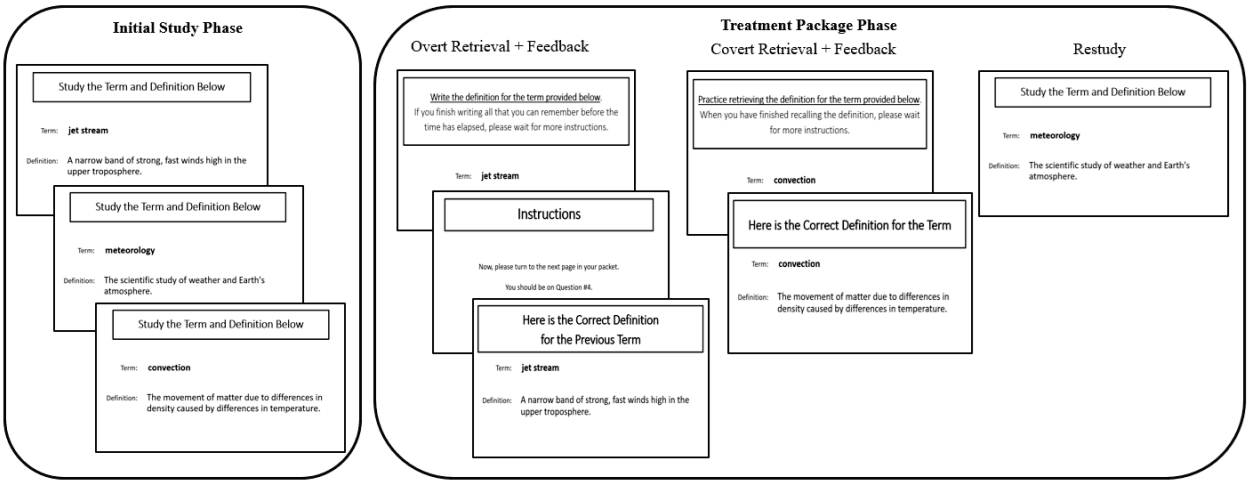


Figure 2. Initial study and treatment package procedures. The initial study phase involved studying 15 terms and their definitions. The treatment package phase involved studying the terms from the initial study phase in one of three ways: overt retrieval with feedback, covert retrieval, or restudy.

Table 1. Number of students in each counterbalanced condition during the learning sessions (i.e., Sessions 1–5). O = Overt; R = Restudy; C = Covert.

	Treatment Package Order					
Session	**O, R, C**	**O, C, R**	**R, O, C**	**R, C, O**	**C, O, R**	**C, R, O**
1	12	18	20	0	0	15
2	0	12	0	15	18	20
3	15	0	12	18	20	0
4	20	0	18	12	15	0
5	0	15	0	20	12	18
Total	47	45	50	65	65	53

During each session, there was a block of 5 terms in the restudy treatment package, 5 terms in the overt retrieval treatment package, and 5 terms in the covert retrieval treatment package. The order of term and definition pairs within each block was randomized per session. To illustrate, convection was learned with the covert retrieval treatment package by the 12 students in the 9 a.m. class. In session 1, it was the 5th term in the third block (O,R,C order); in session 2, it was the 4th term in the second block (O,C,R order); in session 3, it was the 3rd term in the third block (R,O,C order); in session 4, it was the 1st term in the second block (R,C,O order); and in session 5, it was the 2nd term in the first block (C,O,R order). Thus, across classes, each term and definition were learned using each treatment package, with one exception. Due to experimenter error, the term meteorology was not assigned to the overt treatment package. A limitation of this approach is that the assignment of term and definition pairs to each treatment package across sessions was unbalanced (see Table 2). This occurred for two reasons. First, the number of students enrolled differed between classes (*n*s = 12, 15, 18, 20). Second, as noted above, the overt retrieval treatment package was never offered for one term (meteorology), which led to the restudy treatment package being offered in three classes. As a result, to-be-learned content was confounded with classes, a concern that we revisit in the General Discussion.

Table 2. Number of students who learned via each treatment package for each science term. * number of students reflects experimenter error in procedure.

	Treatment Package Type		
Term	**Overt**	**Covert**	**Restudy**
Cold Front	35	18	12
Convection	18	32	15
Convection Current	20	18	27
Coriolis Effect	18	32	15
Deep Current	32	18	15
Global Winds	18	27	20
Jet Stream	27	20	18
Local Winds	18	35	12
Meteorology	0 *	15	50 *
Ocean Current	32	15	18
Surface Current	15	32	18
Upwelling	30	15	20
Warm Front	32	18	15
Weather	15	12	38
Weather Map	15	18	32

For the overt retrieval treatment package, students were given a handout with space to write the definition of the terms. Each term appeared on an individual page, and each page had space for students to write the definition for the corresponding term. Students received the following instructions:

> "You are now going to take a **practice test on some of the terms and definitions**. To do so, you will practice retrieving the definition to each term in preparation for the final memory test.
>
> You will be presented with each term one-at-a-time. For this test, your task is to **write down the definition to each on the paper provided**. You will be given 30 s to write each definition.
>
> **IMPORTANT: You should do your best to recall and write down as much of the definition as you can.** Think about this like short answer questions on a test, in which you are asked to write down the whole definition for a term. The important thing is to **really try to recall and write down each definition.** When you have finished writing the definition for the term, we will provide you with **feedback as to the correct definition for it**, which we will provide on the screen and read aloud."

Then, a term was presented, and students wrote its definition. Using the time feature for slide transitions in PowerPoint, students were allowed 30 s for overt retrieval, but it was immediately apparent that this did not give them adequate time to write the definition for each term. Thus, after the first session, the time for overt retrieval was extended to approximately 90 s per term for students in all classes. Following overt retrieval, students turned the page to prevent them from editing their responses, and they received feedback for the term. Specifically, the term and full definition were presented. Students turned in their handouts after they had written the definition and received feedback for the 5 terms assigned to overt retrieval practice.

Unlike the overt retrieval treatment, students did not write down their responses when presented with term in the absence of the definition for the covert retrieval treatment. For the covert retrieval treatment package, consistent with prior research (e.g., Tauber et al., 2018) students received the following instructions:

> "You are going to take a **practice test on some of the terms and definitions**. To do so, you will practice retrieving the definition to each term in preparation for the final memory test.
>
> You will be presented with each term one-at-a-time. For this test, your task is to **silently retrieve the definition to each term**. You will be given 30 s to think of the definition for each.
>
> **IMPORTANT: You should do your best to recall as much of the definitions as you can.** Think about this as if you and a friend were quizzing each other on terms for an upcoming test in your class—if your friend gave you a term, you'd try to come up with the entire definition. Or, think about it like a practice test, in which you were asked to retrieve the entire definition for each term. In both cases, you would try to retrieve the entire definition for each term. That is what we'd like you to do on each trial here.
>
> The important thing is to **really try to recall each definition**. On each trial, you should not just read the term and assume that you know it because it seems familiar. Rather, use the term as a cue to try to silently recall the definition from memory. When you have finished retrieving the definition for the term, we will provide you with **feedback as to the correct definition for it**, which we will provide on the screen and read aloud.
>
> Remember, you are learning the definitions for each term in preparation for a future memory test. On that test you will be provided with the terms, and you will be required to write the full and correct definition for each. Do you have any questions before we begin?"

Each term was then presented alone for 30 s, which was ensured by using the time feature for slide transitions in PowerPoint, and students used the time to covertly retrieve the definition for each term. The researchers asked the students if more time was needed for covert retrieval, and because no students expressed the need for additional time, the time remained the same across sessions. Students were not allowed to write anything during covert retrieval. Students received corrective feedback that was identical to that of overt retrieval.

For the restudy treatment package, students received the following instructions:

> "You are going to **restudy some of the terms and definitions**. To do so, you will study each term and definition in preparation for the final memory test. The terms will be presented one-at-a-time on the screen and will be read aloud. You will be given 30 s to study each definition. Do you have any questions before we begin?"

The PowerPoint slide transitions were set so that each term and definition was presented alone for 30 s, and students used the time to study each. Students were not allowed to write anything during restudy. The researchers asked the students if more time was needed for restudy, and because no students expressed the need for additional time, the time remained the same across sessions. Thus, the restudy treatment package differed from both overt and covert retrieval because the term and definition were always presented together, which prevented students from needing to retrieve anything from memory to define them.

The session 2 procedure was identical to that of session 1. Specifically, students had an initial study phase of all 15 terms and their associated definitions, which were presented one at a time for 15 s each using PowerPoint slide transitions. The order of the items was randomized anew. Then, students learned 5 terms and definitions with the covert retrieval

treatment package, 5 terms and definitions with the overt retrieval treatment package, and 5 terms and definitions with the restudy treatment package. In sessions 3–5, the initial study phase was eliminated. Otherwise, the procedure was identical to that of sessions 1 and 2.

Session 6 occurred two days after session 5. Students received instructions that they would take a short-answer test, a multiple-choice test, and complete a demographic form. All tasks were self-paced. Students in each class received the same instructions and completed the tasks in the same order. The following instructions were given before beginning the short-answer test:

> "First, you are going to take a **final short-answer test** over the terms and definitions that you have studied last week. For each question, your task is to write the full and correct definition for each term. So, please do your best to recall as much as you can for each question!
>
> **Important:** you must do your best to **answer EVERY question** on this test. You cannot leave any blank. Once you have your paper, you are free to begin! If you have any questions, simply raise your hand and a researcher will come to you."

On the short-answer test, students were given a paper packet with the 15 key terms with space for students to write the definition. The order of terms in the packet was randomized per class, and students could complete the test by writing definitions to the terms in any order desired (e.g., they could skip terms and then return to them). A researcher checked to make sure that students had attempted to define each term. If any definitions were left blank, the test was returned, and the student was encouraged to try their best to define all terms. No feedback was provided on the short-answer test. Students then received the following instructions for the multiple-choice test:

> "Next, you are going to take a **final multiple-choice test** over the terms and definitions that you have studied last week. For each question, your task is to select the definition that best corresponds to the term.
>
> **Important:** you must do your best to **answer EVERY question** on this test. You cannot leave any blank. Once you have your paper, you are free to begin! If you have any questions, simply raise your hand and a researcher will come to you."

For the multiple-choice test, students were provided with a paper packet with multiple questions on each page. The test consisted of 15 four-alternative forced-choice questions (see materials posted online). Thus, for each question, a term was provided (e.g., *What is the Coriolis Effect?*) with four options from which to select. Similar to the short-answer test, for the multiple-choice test, questions order was randomized per class, and the options for each question were presented in the same fixed order for all students. Students could complete the test by answering the questions in any order (e.g., they could skip questions and return to them later). Feedback was not provided on the multiple-choice test. Researchers checked that all questions were answered before allowing the student to move to the next task. If there were questions that had not been answered, the test was returned to the student to complete. Last, students and the teacher completed a demographic form, were debriefed, and were compensated. Students were also provided with a debriefing statement (in English or Spanish) to bring home to their guardians.

The teacher reported that during the experiment (all 6 sessions) she did not do anything differently with her four classes that deviated from normal classroom activities. On 3 days for all classes, at least one term that was used in the experiment was used during class time outside of the context of the experiment. For one of these days, students completed

a textbook reading that covered most of the terms being taught during the experiment. Students had no other assignments, tests, or class activities associated with the terms on Earth's oceans and atmospheres.

2.5. Data Scoring

Each definition was composed of idea units. For example, the definition for the term *jet stream* contained 2 idea units: (1) *A narrow band of strong, fast winds* (2) *high in the upper troposphere*. The number of idea units ranged from 1 to 5 per definition (see the online supplement for the scoring rubric). Two independent raters scored all responses for overt retrieval practice from sessions 1–5 and for the short-answer test from session 6. For each idea unit, trained raters assigned 0 (no credit), 0.5 (partial credit), or 1 (full credit), depending on the completeness and correctness of each student's response. The mean percent correct for each term was calculated separately for each student and for each rater. For instance, the response "*the movement of air in the upper troposphere*" received partial credit (0.5) for unit 1 and full credit (1) for unit 2 from both independent raters, resulting in 1.5 total credits out of the 2 total credits possible for this concept, resulting in 75% correct for the concept jet stream this participant. For 71.4% of responses (2345 total), the raters' scores were in agreement. Interrater reliability was acceptable to high (Pearson's $r = 0.640$ to $r = 0.985$) and significant in all instances ($ps \leq 0.001$; see https://osf.io/p4cq8, accessed on 16 June 2025 for all correlations). Even so, a third independent rater viewed and resolved all conflicts. Thus, all responses were scored by either 2 or 3 independent raters. The resulting scores were then used to calculate the percentage of idea units correctly retrieved during practice and on the short-answer test by summing the idea units earned for each term separately for each participant, dividing by the by the total idea units possible for that term, and multiplying by 100. Aggregate averages were then created by summing the percentage of idea units for each term and dividing by the number of students who contributed responses on the short-answer test to create the average for each term. Aggregate averages were created using the same steps for performance during overt retrieval practice.

3. Results

The outcomes of primary interest were students' performance on the final short-answer and multiple-choice tests, so they are presented first. Next, we present students' performance for overt retrieval during practice sessions. Because class sample sizes were low (i.e., $n = 12$ to $n = 20$), our goal was not to evaluate between-classroom effects, and the power analysis was conducted for the within-participant manipulation rather than the between-participant class variable, we focused analyses on the within-participant manipulation of primary interest. Raw data and analyses by class can be found in the online supplement (https://osf.io/p4cq8, accessed on 16 June 2025).

3.1. Analytic Plan

We used the Bonferroni correction for all pairwise comparisons. The Bonferroni correction adjusts the observed *p*-value to account for multiple comparisons of the same data, which reduced the chance of committing a Type 1 error. We also report Bayes factors for all analyses (BF) (Krushcke, 2013), and we used JASP statistical software version 0.18.3 (JASP Team, 2024) to conduct all Bayes analyses (e.g., Faulkenberry et al., 2020; van Ravenzwaaij & Etz, 2021). We used a Cauchy prior distribution with the JZS prior (r = 0.707). In contrast to traditional null hypothesis significance tests, Bayes factors allow for the quantification of evidence in support of both the null and alternative hypotheses. BF_{01} is a ratio that indicates the degree to which the null hypothesis predicts the data better than the alternative hypothesis. A BF_{01} value of 1 indicates that the null and alternative hypothesis

are equally probably. A BF_{01} value less than 1 provides evidence that more strongly favors the alternative hypothesis, whereas a BF_{01} value greater than 1 provides evidence that more strongly favors the null hypothesis. In the case of all non-significant analyses, we report BF_{01}. For ease of interpretation, for all significant effects, we report BF_{10}, which is simply the reciprocal of BF_{01}, and it indicates the likelihood that the alternative hypothesis better predicts the data compared to the null hypothesis. BF_{10} is interpreted similarly to BF_{01} (i.e., $BF_{10} = 1$ indicates equal support for the null and alternative hypotheses; $BF_{10} < 1$ indicates stronger support for the null over the alternative hypothesis; $BF_{10} > 1$ indicates stronger support for the alternative over the null hypothesis).

3.2. Percent Correct on the Final Short-Answer Test

Students' performance on the short-answer test tended to benefit most from the overt retrieval treatment package, and the impact of the covert retrieval treatment was similar to that of the restudy treatment (see Figure 3). In support of these conclusions, a 3-level (treatment package: overt, covert, or restudy) within-participant effect analysis of variance (ANOVA) was conducted on the percent correct on the final short-answer test. The analysis revealed that treatment package significantly impacted the percent correct on the final short-answer test, $F(2, 126) = 12.76$, $p < 0.001$, $\eta_p^2 = 0.17$, $BF_{10} = 1960.10$. The overt retrieval treatment ($M = 39.19$, $SE = 2.88$) led to significantly higher performance on the short-answer test relative to the covert retrieval treatment ($M = 27.16$, $SE = 2.44$) ($p < 0.001$, $d = 0.56$, $BF_{10} = 354.65$) and restudy treatment ($M = 29.4$, $SE = 2.17$) ($p < 0.001$, $d = 0.50$, $BF_{10} = 161.08$). There was no significant difference between the covert and restudy treatments on the short-answer test ($p = 0.427$, $BF_{01} = 5.38$).

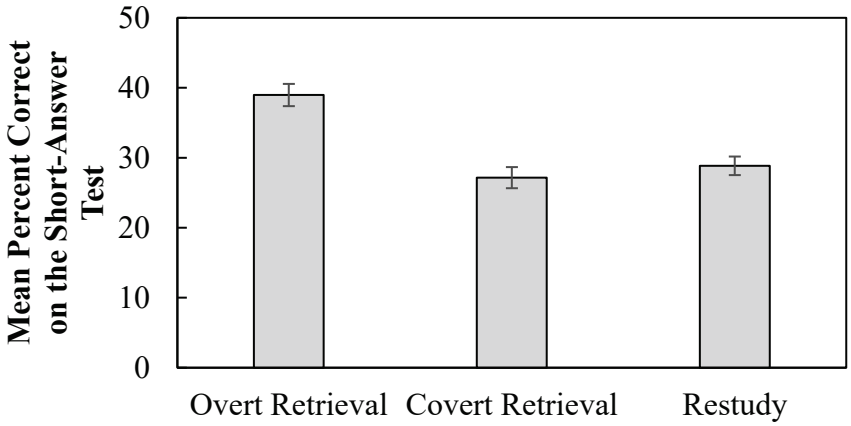


Figure 3. Mean percent correct on the short-answer test. Error bars reflect within-participant standard error of the mean (Loftus & Masson, 1994).

Due to materials being confounded with classes, we also calculated average performance at the idea-unit level of each term separately for each treatment package. The term meteorology was not included in this data treatment to avoid confounding the calculation due to missing values. Consistent with the primary analysis above, a 3-level (treatment package: overt, covert, or restudy) repeated measures ANOVA revealed that there was a significant main effect of treatment package ($F(2,86) = 8.84$, $p < 0.001$, $\eta_p^2 = 0.17$, $BF_{10} = 78.26$). The pairwise comparisons confirmed that recall was significantly higher in the overt retrieval treatment package ($M = 0.35$, $SE = 0.03$) than in the covert retrieval ($M = 0.27$, $SE = 0.03$) ($p = 0.001$, $d = 0.35$, $BF_{10} = 223.91$) and restudy ($M = 0.27$, $SE = 0.03$) ($p = 0.001$, $d = 0.33$, $BF_{10} = 16.99$) treatment packages. Recall did not differ between covert and restudy treatments ($p = 1.00$, $BF_{01} = 6.12$).

For exploratory purposes, we also report analyses at the item level, given that the terms were unbalanced across treatment packages. The mean proportion correct per term and treatment package is reported in Table 3.

Table 3. Mean percent correct on the final short-answer test for each science term and treatment package. Standard errors are provided in parentheses. Terms are presented with the number of idea units for each. Due to experimenter error, the term meteorology was not learned using the overt retrieval treatment package. * significant impact of treatment package + non-significant trend of treatment package.

		Treatment Package Type			
Term	**Idea Units**	**Overt**	**Covert**	**Restudy**	**Overall**
Cold Front	4	42.86 (4.98)	35.29 (7.29)	52.08 (10.64)	42.58 (3.88)
Convection	3	11.76 (3.12)	8.85 (2.59)	8.89 (2.75)	9.64 (1.65)
Convection Current	3	13.33 (3.75)	9.80 (3.52)	17.28 (4.02)	14.06 (2.26)
Coriolis Effect +	4	15.28 (3.58)	5.08 (2.62)	8.33 (2.64)	8.65 (1.80)
Deep Current *	3	58.85 (6.03)	37.25 (6.31)	43.33 (6.04)	49.48 (3.87)
Global Winds	1	52.78 (8.55)	62.96 (5.72)	42.50 (7.50)	53.85 (4.15)
Jet Stream +	2	40.74 (7.45)	18.75 (6.25)	20.83 (7.34)	28.46 (4.31)
Local Winds	2	13.24 (3.79)	7.14 (2.19)	16.67 (5.62)	10.55 (1.91)
Meteorology	3		47.78 (8.74)	38.67 (4.22)	40.77 (3.82)
Ocean Current	2	76.56 (5.01)	66.67 (7.97)	65.28 (8.37)	71.15 (3.85)
Surface Current	3	32.22 (7.71)	24.48 (3.89)	16.67 (5.35)	24.22 (3.4)
Upwelling	3	23.56 (6.22)	34.44 (10.72)	8.33 (5.05)	21.35 (4.21)
Warm Front	4	42.97 (5.83)	42.65 (9.10)	50.00 (8.80)	44.53 (4.26)
Weather	5	47.33 (5.11)	34.17 (7.33)	23.51 (4.03)	31.09 (3.17)
Weather Map	5	32.67 (5.97)	21.76 (5.51)	23.13 (3.40)	25.00 (2.65)

3.3. Percent Correct on the Final Multiple-Choice Test

As evident from Figure 4, the overt retrieval treatment was associated with increased performance on the final multiple-choice test, whereas the covert retrieval treatment was not. In support of this conclusion, a 3-level (treatment package: overt, covert, or restudy) within-participant ANOVA was conducted on the percent correct on the multiple-choice test. Treatment package significantly impacted performance on the multiple-choice test ($F(2, 122) = 4.86$, $p = 0.009$, $\eta_p^2 = 0.07$, $BF_{10} = 1.69$). Performance was significantly higher for definitions that received the overt retrieval treatment ($M = 0.62$, $SE = 0.04$) relative to definitions that received the covert retrieval treatment ($M = 0.50$, $SE = 0.03$) ($p = 0.007$, $d = 0.38$, $BF_{10} = 4.19$). However, there was no significant difference in performance between definitions that received the overt retrieval and restudy treatments ($M = 0.55$, $SE = 0.03$) ($p = 0.17$, $BF_{01} = 1.99$).

We also explored multiple-choice outcomes at the item level. The mean proportion correct per term and treatment package is reported in Table 4. We computed a repeated-measures ANOVA on these means to evaluate whether treatment package affected multiple choice performance at the item level. The term meteorology was not included in this analysis to avoid confounding the calculation due to missing values. The outcomes revealed a non-significant effect of treatment package ($F(2, 26) = 1.95$, $p = 0.169$, $\eta_p^2 = 0.13$, $BF_{01} = 1.56$). Even so, one limitation of this analysis is that it was underpowered given that there are only 14 data points per treatment package. Consistent with this, the outcome of the Baye's analysis did not provide strong evidence for or against either the null or alternative hypothesis. Numerically, the outcomes were consistent with the above analysis such that performance was higher for the overt retrieval treatment ($M = 61.67$, $SE = 5.22$) than for either the covert retrieval treatment ($M = 51.48$, $SE = 4.83$) or the restudy treatment ($M = 54.48$, $SE = 5.52$).

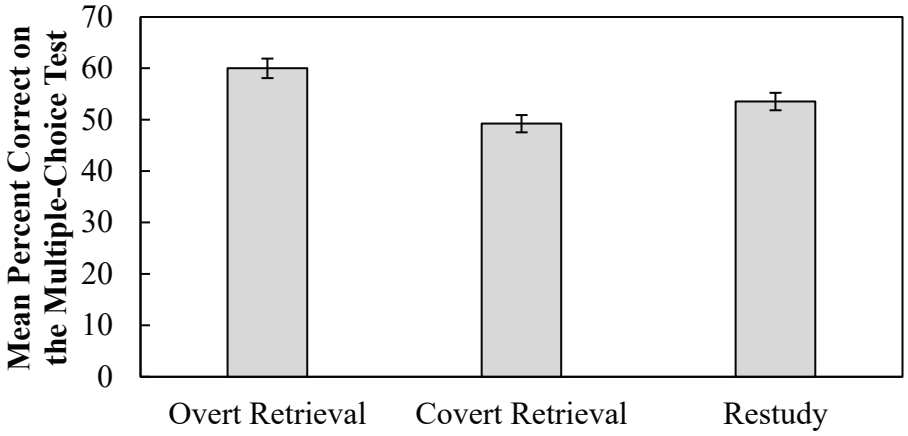


Figure 4. Mean percent correct on the multiple-choice test. Error bars reflect within-participant standard error of the mean (Loftus & Masson, 1994).

Table 4. Mean percent correct on the multiple-choice test for each science term and treatment package. Standard errors are provided in parentheses. Terms are presented with the number of idea units for each. Due to experimenter error, the term meteorology was not learned using the overt retrieval treatment package. * significant impact of treatment package + non-significant trend of treatment package.

		Treatment Package Type			
Term	**Idea Units**	**Overt**	**Covert**	**Restudy**	**Overall**
Cold Front	4	65.71 (8.14)	66.67 (11.43)	75.00 (13.06)	67.69 (5.85)
Convection	3	44.44 (12.05)	31.25 (8.32)	46.67 (13.33)	38.46 (6.08)
Convection Current	3	25.00 (9.93)	33.33 (11.43)	29.63 (8.96)	29.23 (5.69)
Coriolis Effect *	4	61.11 (11.82)	25.00 (7.78)	60.00 (13.09)	43.08 (6.19)
Deep Current	3	65.63 (8.53)	61.11 (11.82)	86.67 (9.09)	69.23 (5.77)
Global Winds +	1	44.44 (12.05)	55.56 (9.75)	20.00 (9.18)	41.54 (6.16)
Jet Stream	2	74.07 (8.59)	55.00 (11.41)	66.67 (11.43)	66.15 (5.91)
Local Winds	2	44.44 (12.05)	34.29 (8.14)	58.33 (14.86)	41.54 (6.16)
Meteorology	3		80.00 (10.69)	62.00 (6.93)	66.15 (5.91)
Ocean Current +	2	83.33 (11.24)	48.57 (8.57)	72.22 (10.86)	61.54 (6.08)
Surface Current	3	66.67 (12.60)	43.75 (8.91)	50.00 (12.13)	50.77 (6.25)
Upwelling	3	43.33 (9.20)	46.67 (13.33)	20.00 (9.18)	36.92 (6.03)
Warm Front	4	71.89 (8.08)	55.56 (12.05)	66.67 (12.60)	66.15 (5.91)
Weather *	5	73.33 (11.82)	91.67 (8.33)	42.11 (8.12)	58.46 (6.16)
Weather Map +	5	100 (00)	72.22 (10.86)	68.75 (8.32)	76.92 (5.27)

3.4. Percent Correct During Overt Retrieval Practice

Percent correct during overt retrieval practice could only be assessed for terms and definitions assigned to the overt retrieval treatment because this was the only condition that required a response during learning. Students' overt retrieval practice tended to improve across learning sessions (see Figure 5). In support of this conclusion, a 5-level (learning session: 1, 2, 3, 4, 5) within-participant ANOVA was conducted. Overt retrieval practice significantly differed between the sessions ($F(4, 172) = 8.32$, $p < 0.001$, $\eta_p^2 = 0.16$, $BF_{10} = 1292.59$). Session 5 overt retrieval practice was significantly higher relative to sessions 1–3 (ps ≤ 0.02, ds 0.28–0.54, $BFs_{10} > 9.60$). No other comparisons were significant (ps ≥ 0.07, $BFs_{01} > 0.19$).

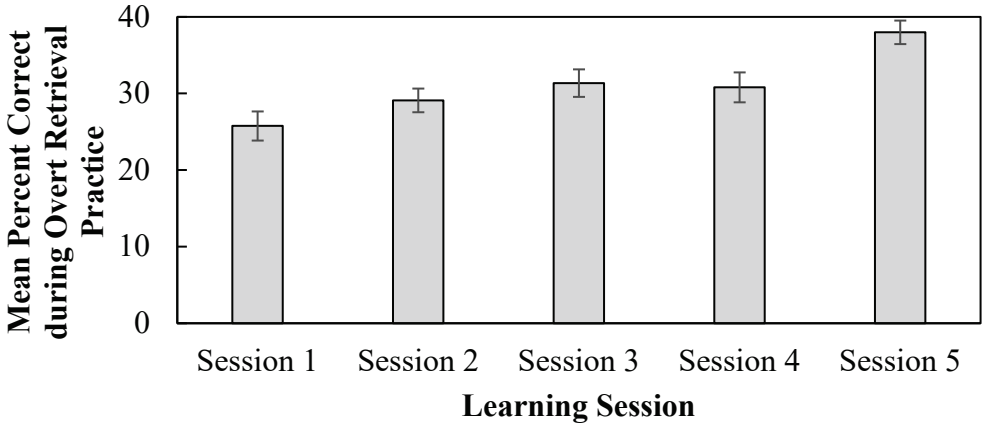


Figure 5. Mean percent correct during overt retrieval across learning sessions. Error bars reflect within-participant standard error of the mean (Loftus & Masson, 1994).

4. Discussion

This experiment is the first to evaluate the impact of a covert retrieval treatment package on adolescents' learning, and we explored this issue by using class concepts and evaluating students' learning in middle school classrooms. Of primary interest, the covert retrieval treatment during multiple study sessions was less effective for adolescents' long-term learning relative to the overt retrieval treatment. This inefficacy was found for both the short-answer and the multiple-choice tests, although specific terms may have contributed to the effect on the multiple-choice test. Consistent with these outcomes, students and teachers rate covert retrieval to be less effective than overt retrieval, and teachers indicate being less likely to recommend covert than overt retrieval to students (Witherby et al., 2025). Adolescents' learning did not substantially differ depending on whether they received the covert retrieval or restudy treatment package. These outcomes are consistent with prior work demonstrating that response modality (covert versus overt retrieval) can moderate the impact of retrieval practice on students' learning (Jönsson et al., 2014; Abel & Roediger, 2018; Kubik et al., 2020, 2022; Tauber et al., 2015, 2018; Ariel et al., 2021). As important, the reported outcomes advance the literature by establishing a treatment package effect for adolescents engaging in retrieval practice in educational contexts.

In contrast to the covert retrieval treatment, the overt retrieval treatment benefited students' long-term learning. As expected, the overt retrieval treatment package led to higher average scores on the short-answer test relative to restudying the terms and definitions. Although the overt retrieval treatment did not statistically benefit students' learning relative to the restudy treatment on the multiple-choice test, a small numerical advantage (overt retrieval treatment $M = 0.62$, $SE = 0.04$; restudy treatment $M = 0.55$, $SE = 0.03$) was observed. These outcomes are consistent with prior research (e.g., Darabi et al., 2007; Kang et al., 2011; Pan et al., 2016, 2019; for reviews, see Agarwal et al., 2012, 2021; Moreira et al., 2019; Roediger & Karpicke, 2006b).

Why can response mode impact the effectiveness of treatment package? There are likely two contributing factors. According to the retrieval dynamics hypothesis (Tauber et al., 2018), when learning information with multiple components, students face a high retrieval burden, which is not well supported by covert retrieval. Consider for instance that to correctly retrieve the definition for the term *cold front*, students must retrieve four complex units of information—(1) *the zone separating two air masses* (2) *of which the cooler, dense mass*, (3) *is advancing and replacing*, (4) *the warmer mass*. When covertly retrieving this definition from memory, students may halt retrieval prematurely due to familiarity with the concept or incorrect beliefs that information is well learned. By doing so, exhaustive

retrieval is avoided, which decreases the effectiveness of retrieval practice. By contrast, because overt retrieval requires a written, typed, or spoken response, it is clearer when only partial (or no) information is generated. This likely encourages students to attempt full retrieval more often, which increases the effectiveness of retrieval practice. When the retrieval burden is low because students are learning less complex information, as used in some prior research, covert retrieval can be as effective as overt retrieval (Higham et al., 2023; Putnam & Roediger, 2013; Smith et al., 2013; Sundqvist et al., 2017). This is positive because covert retrieval can be relatively quick and easy to implement in classrooms. Teachers can construct retrieval practice questions on the fly and provide immediate feedback to their students. By contrast, methods to implement overt retrieval require more preparation for teachers (e.g., constructing quizzes before class) and can require additional course equipment (e.g., iClickers). Overt retrieval also requires that time be made available so that students can write their answers or say them out loud.

In addition to differences in the dynamics of retrieval between treatment packages, a second key factor to consider is the amount of time students spent on the learning task in each treatment package. The time during learning was held constant for the covert retrieval and restudy treatment packages and no students expressed needing more time to recall the definition or restudy it, which is consistent with college-aged students' self-paced measures of reaction time during covert retrieval and restudy (Tauber et al., 2018). However, participants took longer to complete their overt retrieval responses relative to covertly retrieving definitions or restudying them. This unequal study time between treatment packages is important for overt retrieval outcomes. Specifically, it is likely that some portion of the benefit of the overt retrieval treatment package is attributable to increased time with the materials. This is a natural occurrence for the overt retrieval treatment package because it takes more time to generate a manual response relative to a mental response. As important, this experiment provides an index of the time needed by adolescents to use retrieval practice strategies in real educational settings. We think it unlikely that exposure time alone drives the benefits of the overt retrieval treatment because ample prior work has established that the quality of the learning strategy employed can overshadow simple exposure time.

Our outcomes suggest that response mode (overt vs. covert retrieval) impacts the effectiveness of students' learning due to the dynamics of retrieval and the amount of time necessary to complete learning the learning task. The amount of time required to practice retrieval will depend on the response mode and complexity of the to-be-learned materials, so an important direction for future research will be to disentangle response mode from study time. Doing so will be important for considering not only the effectiveness of each treatment package, but also their efficiency. Students recalled the most with overt retrieval, but overt retrieval also consumed more time relative to covert retrieval and restudy. Increasing exposure time for covert and restudy conditions to match the needed time to generate an overt response would be experimentally simple; however, implementing the learning strategies in this way is unrealistic for an actual classroom context, and this applied focus drove our experimental design. Additional restudy or covert retrieval time beyond what students report as necessary to engage in the task would presumably lead to mind wandering or off-task engagement for many students in a classroom context. Indeed, the classroom teacher and school principal both lamented this as a regular concern in middle school classrooms. An alternative could be to allow adolescents to self-pace their retrieval practice during learning, measure individual reaction times, and establish the relationship between retrieval practice time and final test performance for both covert and overt response modalities. This is a fruitful direction for future research

because it can increase understanding of treatment efficiency—i.e., how much knowledge is gained relative to the amount of time needed to use each strategy during learning.

There are other practical concerns to consider about implementing covert retrieval in middle school classrooms. Covert retrieval does not provide information on the accuracy of students' responses, which makes it challenging for teachers to identify gaps in knowledge and determine if concepts need to be revisited. Unexpected distractions can occur during classes that may be particularly problematic for covert retrieval. Informal notes taken by the research team indicated that several students had their heads down, eyes closed, or were looking around the room and away from the screen during covert retrieval and restudy. In some instances, these behaviors may reflect deep thinking and comprehensive retrieval of definitions, but in others, students may be disengaged. Additional research is needed to systematically probe students for engagement during covert retrieval. To combat disengagement, Sumeracki and Castillo (2022) proposed that teachers combine in-class cold calling with covert retrieval, and an important direction for future research will be to consider this and other approaches for increasing the effectiveness of covert retrieval for supporting adolescents' learning in classroom environments. Another practical direction for future research will be to explore methods to increase the effectiveness of the covert retrieval treatment when learning complex materials. Covert retrieval may necessitate scaffolding for success such as reminding adolescents how many components each definition has (e.g., "remember, the definition for this term has 4 key components and you should try to recall them all") or teaching students strategies for updating their knowledge when feedback is provided (e.g., "locate all 4 components of the definition in the feedback. Which did you remember and which did you forget?").

An important issue for both the educational context and experimental work is to measure student characteristics and explore the degree to which individual differences impact success with each treatment package. For instance, covert retrieval likely involves some ability to rely on an inner monologue, and some students may be better at this than others. Additionally, some students may need more time to engage in covert retrieval relative to others, and not all students will feel comfortable speaking up in class about needing more time to complete a task. Another factor that may be critical is students' prior knowledge or level of expertise when learning new information using each treatment. Addressing these issues will be valuable for increasing understanding of how to best support students' learning.

To contextualize our outcomes, we must note a few important limitations of the reported experiment. The biggest factor limiting aspects of our methodology was conducting this experiment in authentic adolescent science classrooms. Carrying out experimental work in a classroom context is difficult because there are many factors outside of the experimenters' control. Examples include class enrollment numbers, non-random assignment to classes, class time availability, course schedule, and coordinating educational content with the science classes. These conditions make for a noisy assessment of students' learning; however, carrying out this labor-intensive research is critical for investigating the degree to which laboratory effects generalize to a classroom context and a broad range of students (e.g., Agarwal et al., 2021; Schwieren et al., 2017). This step is critical for establishing best recommendations for educators, but reduced methodological control can make it challenging to draw firm conclusions from obtained classroom outcomes.

Given the classroom nature of this work, a limitation of our research is that counterbalancing of treatment packages and terms across classes and sessions was unbalanced and therefore terms were confounded with classes. This was driven by variance in class sizes, having access to only four science classes, and for one term, experimenter error (see procedure for details). Relevant to this issue, even though the course teacher verified that

students had not been taught the STEM materials used in our experiment, we did not measure students' prior knowledge of the concepts. Reported outcomes should be interpreted as directly relevant to the materials used with the degree to which they generalize to other materials as an open question for subsequent work. We recommend laboratory research to replicate this applied experiment adopting better counterbalancing as well as classroom research using the same materials as well as other content adolescents are required to learn in their STEM classes. With all these approaches, it will be important to quantify and evaluate the impact of characteristics of the to-be-learned materials (e.g., difficulty, complexity, abstractness) on learning.

We chose to manipulate treatment package using a within-participant design due to practical constraints with the availability of students. This way, all students in all classes received each treatment package, and we avoided assigning treatment packages to individual classes for which students could not be randomly assigned. However, this could be a limitation of the reported experiment because carryover effects may have occurred between treatment packages. For instance, experiencing both overt and covert retrieval could encourage better covert retrieval. Even so, this would reduce the magnitude of our obtained effects suggesting that a between-participant design would likely reveal larger effects. In other words, the impact of treatment package on adolescents' learning may be more substantial if they only get to experience one treatment eliminating the possibility of strategy carry-over.

A limitation inherent to covert retrieval is that it is difficult to establish what students are doing when mentally retrieving information. This same limitation exists for restudy comparisons that do not require students to copy restudied information. However, this issue does not arise for overt retrieval because an objective retrieval response is recorded. These are ongoing challenges for the field. Identifying methodology to address these concerns will be important for refining theory on student learning. One avenue could be to include subjective assessments for students to report the degree to which they tried to retrieve the definition or strategies they used during covert retrieval. Another avenue could be to include a no study control comparison condition in future research to establish the degree to which all treatment packages impact learning relative to doing nothing. This approach would be less useful from an applied perspective, but it would be valuable for establishing the degree to which covert retrieval enhances learning relative to not studying, which provides insight into students' willingness or ability to mentally retrieval information. Another possibility is that feedback may be critical for students' covert retrieval activity. We gave students feedback in all treatment packages to avoid introducing a between-condition confound and, due to the applied nature of this work in authentic middle school classrooms, we prioritized conditions that would be most likely for enhancing student learning (e.g., Bangert-Drowns et al., 1991; Pashler et al., 2005). However, an important path forward for laboratory research will be to investigate the effectiveness of covert retrieval when students learn complex concepts with or without feedback. It is possible that students began to anticipate feedback on covert retrieval trials, which may have led some to disengage in covert retrieval and simply wait for the feedback. This may make covert retrieval more similar to a restudy opportunity in the context of receiving feedback during learning.

5. Conclusions

In conclusion, overt retrieval practice best supported adolescents' learning of concepts in their middle school science classes. Covert retrieval and restudy were typically less effective for students' long-term learning, although we did find some differences between terms and test types. We recommend that teachers and students use overt retrieval when time

is available and the stakes for correct recall are high. Additional research should identify methods to increase the effectiveness of covert retrieval for increasing adolescents' learning.

Author Contributions: Conceptualization, A.E.W., P.E.N. and S.K.T.; methodology, A.E.W., P.E.N. and. S.K.T.; materials A.E.W. and S.K.T.; formal analysis, A.E.W. and S.K.T.; writing—original draft preparation, A.E.W. and S.K.T.; writing—review and editing, P.E.N.; project administration, S.K.T., P.E.N. and A.E.W.; funding acquisition, S.K.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition (Collaborative Grant Number 220020483) and the Andrews Institute of Mathematics & Science Education (2018–2019 Research Grants).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Texas Christian University (protocol code 1710-032-1809 and date of approval 13 August 2019) and the Grants Compliance and Monitoring Board of the Fort Worth Independent School District (protocol code R_SZ9V5oQusGOFyTf and date of approval 3 May 2019).

Informed Consent Statement: Informed consent was obtained from the school principal, science teacher, and guardians (documented in either English or Spanish). All adolescents assented to take part in the experiment.

Data Availability Statement: All materials, raw data, and class-level analyses have been uploaded to the Open Science Framework and can be freely accessed at https://osf.io/p4cq8 (accessed on 16 June 2025). The study's design and analyses were not pre-registered.

Acknowledgments: We would like to thank Ashlyn Holliday and Megan Pleshek for their assistance with data collection and Payton Adeylott, Kahli Hugley, Truly Juneau, Katherine Neely, and Rachel Perez for their assistance with scoring.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Abel, M., & Roediger, H. L. (2018). The testing effect in a social setting: Does retrieval practice benefit a listener? *Journal of Experimental Psychology: Applied*, *24*(3), 347–359. [CrossRef]

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. [CrossRef]

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review*, *24*(3), 437–448. [CrossRef]

Agarwal, P. K., D'Antonio, L., Roediger, H. L., McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition*, *3*(3), 131–139. [CrossRef]

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*, *33*(4), 1409–1453. [CrossRef]

Ariel, R., Karpicke, J. D., Witherby, A. E., & Tauber, S. K. (2021). Do judgments of learning directly enhance learning of educational materials? *Educational Psychology Review*, *33*(2), 693–712. [CrossRef]

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, A. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*, 213–238. [CrossRef]

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, *1*(9), 496–511. [CrossRef]

Coyne, J. H., Borg, J. M., DeLuca, J., Glass, L., & Sumowski, J. F. (2015). Retrieval practice as an effective memory strategy in children and adolescents with traumatic brain injury. *Archives of Physical Medicine and Rehabilitation*, *96*(4), 742–745. [CrossRef]

Darabi, A. A., Nelson, D. W., & Palanki, S. (2007). Acquisition of troubleshooting skills in computer simulation: Worked example vs. conventional problem solving instructional strategies. *Computers in Human Behavior*, *23*(4), 1809–1819. [CrossRef]

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. [CrossRef]

Faulkenberry, T. J., Ly, A., & Wagenmakers, E. (2020). Bayesian inference in numerical cognition: A tutorial using JASP. *Journal of Numerical Cognition*, *6*, 231–259. [CrossRef]

Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology*, *21*(4), 499–526. [CrossRef]

Higham, P. A., Fastrich, G. M., Potts, R., Murayama, K., Pickering, J. S., & Hadwin, J. A. (2023). Spaced retrieval practice: Can restudying trump retrieval? *Educational Psychology Review*, *35*, 98. [CrossRef]

JASP Team. (2024). *JASP* (Version 0.18.3, Computer software). JASP Team, University of Amsterdam.

Jönsson, F. U., Kubik, V., Sundqvist, M. L., Todorov, I., & Jonsson, B. (2014). How crucial is the response format for the testing effect? *Psychological Research*, *78*, 623–633. [CrossRef]

Kang, S. H., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, *18*(5), 998–1005. [CrossRef]

Krushcke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*, 573–603. [CrossRef]

Kubik, V., Jönsson, F. U., de Jonge, M., & Arshamian, A. (2020). Putting action into testing: Enhanced retrieval benefits long-term retention more than covert retrieval. *Quarterly Journal of Experimental Psychology*, *73*(12), 2093–2105. [CrossRef] [PubMed]

Kubik, V., Koslowski, K., Schubert, T., & Aslan, A. (2022). Metacognitive judgments can potentiate new learning: The role of covert retrieval. *Metacognition and Learning*, *17*(3), 1057–1077. [CrossRef]

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490. [CrossRef]

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414. [CrossRef]

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513. [CrossRef]

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*(3), 360–372. [CrossRef]

McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology*, *72*(1), 609–633. [CrossRef] [PubMed]

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roedgier, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*(1), 3–21. [CrossRef]

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: A review of applied research. *Frontiers in Education*, *4*, 1–16. [CrossRef]

Pan, S. C., Arpita, G., & Rickard, T. C. (2016). Testing with feedback yields potent, but piecewise, learning of history and biology facts. *Journal of Educational Psychology*, *108*(4), 563–575. [CrossRef]

Pan, S. C., Hutter, S., D'Andrea, D., Unwalla, D., & Rickard, T. C. (2019). In search of transfer following cued recall practice: The case of process-based biology concepts. *Applied Cognitive Psychology*, *33*(4), 629–645. [CrossRef]

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 3–8. [CrossRef] [PubMed]

Putnam, A. L., & Roediger, H. L. (2013). Does response mode affect amount recalled or the magnitude of the testing effect? *Memory & Cognition*, *41*(1), 36–48. [CrossRef]

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin*, *137*(1), 131–148. [CrossRef] [PubMed]

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395. [CrossRef] [PubMed]

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. [CrossRef] [PubMed]

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. [CrossRef]

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. [CrossRef] [PubMed]

Schwieren, J., Barengerg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching*, *16*(2), 179–196. [CrossRef]

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1712–1725. [CrossRef]

Sumeracki, M. A., & Castillo, J. (2022). Covert and overt retrieval practice in the classroom. *Translational Issues in Psychological Science*, *8*(2), 282–293. [CrossRef]

Sundqvist, M. L., Mäntylä, T., & Jönsson, F. U. (2017). Assessing boundary conditions of the testing effect: On the relative efficacy of covert vs. overt retrieval. *Frontiers in Psychology*, *8*, 1018–1033. [CrossRef]

Tauber, S. K., Dunlosky, J., & Rawson, K. A. (2015). The influence of retrieval practice versus delayed judgments of learning on memory: Resolving a memory-metamemory paradox. *Experimental Psychology*, *62*(4), 254–263. [CrossRef]

Tauber, S. K., Witherby, A. E., Dunlosky, J., Rawson, K. A., Putnam, A. L., & Roediger, H. L. (2018). Does covert retrieval benefit learning of key-term definitions? *Journal of Applied Research in Memory and Cognition*, *7*(1), 106–115. [CrossRef]

van Ravenzwaaij, D., & Etz, A. (2021). Simulation studies as a tool to understand Bayes factors. *Advances in Methods and Practices in Psychological Science*, *4*, 1–21. [CrossRef]

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate learning of new material. *Psychonomic Bulletin & Review*, *18*(6), 1140–1147. [CrossRef]

Witherby, A. E., Babineau, A. L., & Tauber, S. K. (2025). Students', teachers', and parents' knowledge about and perceptions of learning strategies. *Behavioral Sciences*, *15*(2), 160. [CrossRef] [PubMed]

Witherby, A. E., & Tauber, S. K. (2019). The current status of students' note-taking: Why and how do students take notes? *Journal of Applied Research in Memory and Cognition*, *8*(2), 139–153. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Sorry, Am I Intruding? Comparing Performance and Intrusion Rates for Pretested and Posttested Information

Kelsey K. James [1,*] and Benjamin C. Storm [2]

[1] Department of Psychology, University of Houston Clear Lake, Houston, TX 77058, USA
[2] Department of Psychology, University of California, Santa Cruz, CA 95064, USA; storm@ucsc.edu
* Correspondence: jamesk@uhcl.edu

Abstract

Pretesting and posttesting have long been implemented in classrooms as methods of testing and improving learning. Prior research has been mixed on the relative benefits of pretesting versus posttesting, with some studies finding pretesting to be more beneficial, and others finding posttesting to be more beneficial. True/False testing is a particularly easy-to-implement method and is regularly used in classrooms. However, relatively little is known about how these tests affect learning. Three experiments address the effects of true/false pre- and posttests on learning correct information and intrusion rates of false information. We find consistent benefits of both pretesting and posttesting but significantly higher intrusion rates for posttesting relative to pretesting, a finding that persisted despite inclusion of simple True/False feedback (Experiment 2) and substantive feedback (Experiment 3). Although the difference between pretesting and posttesting intrusion rates was still significant with the addition of substantive feedback, overall intrusion rates were greatly reduced.

Keywords: pretesting; the pretesting effect; testing; the testing effect; true/false testing; learning; errors; the negative suggestion effect

1. Introduction

1.1. Pretesting and Posttesting

The benefits of taking tests after learning are well established and have been demonstrated repeatedly across a multitude of factors, including learning media, testing format, and retention intervals (e.g., Chan & McDermott, 2007; Cranney et al., 2009; Little & Bjork, 2015; Roediger & Karpicke, 2006; Roediger et al., 2011a, 2011b; Rowland, 2014; Spitzer, 1939). In typical studies on the testing effect (hereafter referred to as the posttesting effect), participants study information (e.g., trivia, word pairs, lectures, or passages) and are then tested on a subset of that information before being tested again, often after a delay. Information initially tested via posttest tends to be significantly better recalled at final test than information either not initially tested or simply re-studied, an effect that has been observed in many conditions and with a variety of to-be-learned materials.

Research has shown, however, that taking tests before studying to-be-learned material can also benefit learning, a phenomenon often referred to as the pretesting effect (e.g., Carpenter & Toftness, 2017; Grimaldi & Karpicke, 2012; James & Storm, 2019; Kornell, 2014; Kriechbaum & Bäuml, 2024; Little & Bjork, 2016; Overoye et al., 2021; Pastötter & Bäuml, 2014; Richland et al., 2009). In typical studies on the pretesting effect, participants begin by taking a test on information they have not yet studied. After completing this pretest

(usually with little success), participants are presented with the information needed to answer the questions from the pretest. Similar to posttesting, participants are then given a final test, usually after a delay, and on this final test, participants perform better on pretested information than on non-pretested information.

Although both pretesting and posttesting can be beneficial, their relative effectiveness remains unclear. Some studies have observed a greater benefit of posttesting than pretesting (Geller et al., 2017; Latimier et al., 2019; McDaniel et al., 2011), whereas others have found the opposite (Pan & Sana, 2021). There are compelling reasons to expect posttesting to be more effective. For example, posttesting occurs when learners have already been exposed to the material, a factor which can facilitate retrieval practice performance and make it easier for learners to incorporate new information into memory. In contrast, learners may be unable to answer any questions correctly on a pretest, as they have not yet studied the to-be-learned material. On the other hand, pretesting has the advantage of allowing learners to study the to-be-learned information after taking the pretest, which may focus encoding efforts on that specific information. Indeed, the study phase that follows a pretest can provide direct, powerful feedback, allowing learners to correct errors and solidify knowledge, a feature less likely to take place in the case of posttesting.

1.2. True/False Testing

True/False tests are frequently used in educational settings due to their simplicity and practicality. They are easy to create and grade and can be incorporated into other forms of assessment such as multiple-choice scantron tests. Despite their frequent use, relatively little is known about how they impact learning and whether and to what extent they lead to the kind of pretesting and posttesting benefits described above. There do seem to be benefits, albeit somewhat more mixed compared to what has been observed with other kinds of tests. In one study, for example, Uner et al. (2022) found a posttesting benefit for True/False quizzed items when compared to a re-reading control but not a re-typing control. In another study, Brabec et al. (2021) sought to determine the benefits of True/False posttesting on both tested information and nontested, related information. Interestingly, they found a benefit for tested information when the correct answer on the posttest was "true" and a benefit for nontested information when the correct answer on the posttest was "false." In other recent work on True/False testing, Hildenbrand and Wiley (2025) demonstrated benefits for True/False testing when the final test was an essay (Exp. 1) but not when it was short answer (Exp. 2). Subsequent experiments showed that multiple-choice tests led to better short answer final test performance than True/False tests regardless of feedback (Exp. 3) and text availability (Exp. 4).

An important limitation of True/False testing is that, by the nature of the test, false information is presented to the learners. According to the negative suggestion effect, which was first discussed by Remmers and Remmers (1926), exposure to false information—such as in the context of false items in a True/False test—can lead to worse memory for correct information and a higher likelihood of endorsing false information as true. While Remmers and Remmers did not actually find evidence of a negative suggestion effect in their study, another study by McClusky (1934) found it shortly thereafter, and several other more recent studies found evidence of a negative suggestion effect as well (e.g., Toppino & Brochin, 1989; Toppino & Luipersbeck, 1993). Additionally, related work on multiple-choice testing (Roediger & Marsh, 2005) has shown that multiple-choice tests can lead to a negative suggestion effect, with learners subsequently endorsing incorrect alternatives.

1.3. Pretesting and Posttesting with True/False Tests

The negative suggestion effect from the false items of True/False tests presents a unique opportunity that, to our knowledge, has not been explored in prior research. Specifically, we compare the negative suggestion effect between pretesting and posttesting. Is one kind of test more likely to lead to a negative suggestion effect than the other? We sought to answer this question by measuring the rate of "intrusions" at final test (i.e., the probability of incorrectly recalling a false item that was used in a pre/post TF test).

There are reasons to think that pretesting should lead to more false intrusions than posttesting. Exposure to false information prior to studying true information may leave participants vulnerable to incorporating the false information into existing knowledge structures, thereby making participants more likely to output the false information when tested. When taking a posttest, participants have the benefit of having already studied the true information, putting them in a better position to protect themselves from incorporating the false information. That said, there are reasons to think participants will be less susceptible to false intrusions following pretesting than posttesting. When engaged in pretesting, participants view the information they encounter with more uncertainty, perhaps making them resistant to encoding that information in a way that makes it memorable at final test. Participants may also more effectively discount information that comes to mind when recalled as having been encountered during pretesting, as they would know by virtue of that context that they should be uncertain about its veracity. Finally, participants may activate details from the text when taking a posttest, thereby leading to subsequent confusion at the time of the final test, a factor that would not be relevant in the context of pretesting.

1.4. The Present Experiments

The aim of the present experiments was to examine the benefits and limitations of pretesting and posttesting in the context of True/False questions. More specifically, the goals were to determine (1) the degree to which the benefits of pretesting and posttesting hold up with True/False tests, (2) the relative effectiveness of each method for promoting performance on a final test, and (3) the relative extent to which each test type leads to a negative suggestion effect in the form of intrusions from false information.

To preview our results, no significant differences were found between overall final test performance following pretests and posttests. However, a significantly greater number of false intrusions were found following posttests than pretests. This result was found without feedback (Experiment 1), with simple corrective feedback (just that the answer was true/false; Experiment 2), and with substantive feedback (with the statement repeated with the true information following the guess; Experiment 3), though the difference was much smaller and overall intrusion rates were much lower in the final case.

2. Experiment 1

The primary goal of the first experiment was to compare pretesting and posttesting in the context of true/false test-taking. We sought to compare the benefits of taking such tests not only on their potential to improve final test performance, but on their susceptibility to negative suggestion effects at final test in the form of intrusions from false information presented during pretesting/posttesting. We also examined whether such effects depend on the nature of the true/false items, such as whether the correct answer on a given test was true or false.

2.1. Methods

2.1.1. Participants and Design

A total of 128 undergraduates (*M* age: 20.2 years, range = 18–28; 93 women, 31 men, 2 write-in responses, 2 declined to state) at the University of California, Santa Cruz (UCSC), participated for partial credit in a psychology course as approved by the university's IRB. The sample size was determined prior to beginning data collection. Specifically, we aimed to have enough participants to provide 80% power to observe a relatively small effect size (d = 0.25) when comparing performance and intrusion rates in the Pretest and Posttest conditions. A 2 (Test Type: Pretest vs. Posttest) × 3 (Item Type: True-Tested, False-Tested, Untested) within-subjects design was used. Counterbalancing was implemented such that half the participants received a pretest for the first passage and a posttest for the second passage, and the other half vice versa. Counterbalancing across participants also ensured that each to-be-learned fact was equally likely to be pretested, posttested, or untested and, if tested, whether the statement presented was true or false. For clarity in our analyses, we use the term Question Type to refer to whether a given statement was presented as true or false during the pre/posttest, and Item Type to refer to the learning history of a given fact on the final test (True-Tested, False-Tested, or Untested).

2.1.2. Materials

Two passages were selected from James and Storm (2019): one focused on Joan of Arc (910 words) and the other on Machu Picchu (824 words). For each passage, we created a set of 16 critical facts. To implement our counterbalancing scheme, the 16 facts were distributed across four unique test lists. On a given list, 8 of the 16 facts were assigned to be tested and the other 8 were untested. Of the 8 tested facts, 4 were presented as true statements and 4 were presented as false statements. Across the four lists, each of the 16 critical facts appeared once as a true-tested item, once as a false-tested item, and twice as an untested item, ensuring complete counterbalancing of items across conditions. Each participant was randomly assigned to one of the four lists for their pretest passage and to one of the four lists for their posttest passage. For a given tested item, a participant was shown either the true statement or the false statement, but not both. False items were created with the goal of still making sense within the item with only one word or short phrase changed as compared to the true item.

Final tests consisted of all 16 questions from each passage in cued-recall form. The cued recall items were created either by removing the true/false piece of information (a single word or short phrase) from the initial question and leaving it as a blank or by reframing the initial test item into a question with the answer being the true/false piece of information. For example, the initial T/F item "As part of her mission, Joan of Arc took a vow of chastity; As part of her mission, Joan of Arc took a vow of silence" (true and false, respectively), participants were tested on the cued-recall test with the question "As part of her mission, Joan of Arc took a vow of _____", with the correct answer being "chastity". For the final cued recall test, if a participant answered "chastity" or something with an equivalent meaning, e.g., "celibacy", this would be coded as a "correct" answer, but if they answered with "silence" or something equivalent, this would be coded as an "intrusion" answer. Any other answer or leaving the item blank would be coded as "incorrect". The full set of items, including true/false statements and cued-recall questions, can be found in Appendix A.

2.1.3. Procedure

Participants were tested individually or in groups of up to 3. The experiment took place in person in the lab, with test questions presented over OpenSesame and with the

passages administered on paper. The passages were always administered in the same order for every participant: Joan of Arc then Machu Picchu. Half of the participants received a Pretest for Joan of Arc and a Posttest for Machu Picchu, and half vice versa and participants were randomly assigned to one of these counterbalanced orders. Pretest and Posttest questions were administered one at a time with participants required to select True or False for each item before moving on to the next question. After completing both passages and their respective pre- or posttest, participants were given 5 min to play Sudoku. Then, participants completed a final test including all 32 cued recall items. All Joan of Arc test items were administered first, followed by all Machu Picchu items. Once again, the final test questions were presented one at a time with participants required to write something before moving on to the next question.

Final test answers were coded by the experimenter as either correct, an intrusion, or incorrect. Coding in this experiment, and in all subsequent experiments, was conducted by the first author, who was blind to condition. To be counted as an intrusion, participants needed to output the specific piece of information from the corresponding false item for that question. The coding was done in this way regardless of whether the specific participant had encountered that item as false (i.e., whether or not that participant had actually been exposed to that false information) in order to establish a baseline rate at which participants might output that specific answer. With the example used previously, "As part of her mission, Joan of Arc took a vow of _____", the intrusion answer, "silence", is something that it is not entirely unreasonable to guess based on prior knowledge even if a participant had not encountered it as part of their pre/posttest.

2.2. Results

2.2.1. Pretest and Posttest Performance

We conducted a 2 (Test Type: Pretest vs. Posttest) by 2 (Question Type: True vs. False) repeated-measures ANOVA. As can be seen in the left panel of Figure 1, two main effects were observed such that performance on the posttest ($M = 0.77$, $SD = 0.17$) was significantly better than on the pretest ($M = 0.55$, $SD = 0.19$), $F(1, 127) = 109.10$, $p < 0.001$, $\eta^2_p = 0.46$, and performance on True Questions ($M = 0.76$, $SD = 0.17$) was significantly better than on False Questions ($M = 0.56$, $SD = 0.19$), $F(1, 127) = 83.16$, $p < 0.001$, $\eta^2_p = 0.40$. Moreover, a significant interaction was observed between Test Type and Question Type, $F(1, 127) = 39.52$, $p < 0.001$, $\eta^2_p = 0.24$. Follow-up tests showed that although performance was better on True questions than False questions on both the Pretest, $t(127) = 11.27$, $p < 0.001$, $d = 1.00$, 95% CI [0.78–1.20], $BF_{01} < 0.001$, and the Posttest, $t(127) = 2.55$, $p = 0.01$, $d = 0.23$, 95% CI [0.05–0.40], $BF_{01} = 0.617$, the difference was much greater on the pretest than on the posttest.

The overall gap in performance between performance on the pretest and the posttest is to be expected given that participants had not yet been exposed to the information at the time of the pretest. It is notable, however, that we observed a true response bias in both the posttest and, to a much greater degree, the pretest. It seems that, with these questions, in the absence of additional information, learners err on the side of assuming information is correct. This fits with prior work demonstrating a "veracity bias" (Levine et al., 1999) in which people tend to assume information provided is true.



Figure 1. Pretest and posttest performance across the three experiments as a function of whether the correct answer on a given question was True or False. Error bars represent standard error of the mean.

2.2.2. Final Test Performance

We next compared the magnitude of the benefits of pretesting and posttesting by conducting a 2 (Test Type: Pretest vs. Posttest) by 3 (Item Type: True-Tested vs. False-Tested vs. Untested) repeated-measures ANOVA on final test performance. As can be seen in the left panel of Figure 2, a main effect of Test Type was not observed, $F(1, 127) = 0.49$, $p = 0.47$, $\eta^2_p = 0.00$. Specifically, final test performance did not differ depending on whether participants received a pretest ($M = 0.53$, $SD = 0.22$) or a posttest ($M = 0.55$, $SD = 0.23$). A main effect of Item Type was observed, $F(2, 254) = 58.89$, $p < 0.001$, $\eta^2_p = 0.32$, however, such that True-Tested items ($M = 0.65$, $SD = 0.24$) were recalled better than both False-Tested items ($M = 0.50$, $SD = 0.23$), $t(127) = 7.73$, $p < 0.001$, $d = 0.68$, 95% CI [0.49, 0.88], $BF_{01} < 0.001$, and Untested items ($M = 0.46$, $SD = 0.20$), $t(127) = 10.01$, $p < 0.001$, $d = 0.88$, 95% CI [0.68, 1.09], $BF_{01} < 0.001$. False-Tested items were also recalled better than Untested items, $t(127) = 2.72$, $p = 0.007$, $d = 0.24$, 95% CI [0.06, 0.42], $BF_{01} = 0.405$. Finally, we failed to observe a significant interaction, $F(2, 254) = 0.93$, $p = 0.40$, $\eta^2_p = 0.01$. Thus, we failed to find evidence that participants benefited from pretesting and posttesting differently as a function of Item Type.

2.2.3. Final Test Intrusions

To test our primary hypothesis regarding false intrusions, we conducted a 2 (Test Type: Pretested vs. Posttested) by 3 (Item Type: True-Tested vs. False-Tested vs. Untested) repeated-measures ANOVA on the proportion of intrusions on the final test. This analysis revealed significant main effects of both Item Type, $F(2, 254) = 114.86$, $p < 0.001$, $\eta^2_p = 0.47$, and Test Type, $F(1, 127) = 5.13$, $p = 0.025$, $\eta^2_p = 0.04$, which were in turn qualified by a significant interaction, $F(2, 254) = 6.17$, $p = 0.002$, $\eta^2_p = 0.05$. To decompose this interaction, we directly compared the intrusion rates for the critical False-Tested items across the two test types. As can be seen in the left panel of Figure 3, participants produced significantly more intrusions for items that were false on the Posttest ($M = 0.19$, $SD = 0.21$) than they did for items that were false on the Pretest ($M = 0.13$, $SD = 0.17$), $t(127) = 2.60$, $p = 0.011$, $d = 0.23$, 95% CI [0.05, 0.40], $BF_{01} = 0.389$.

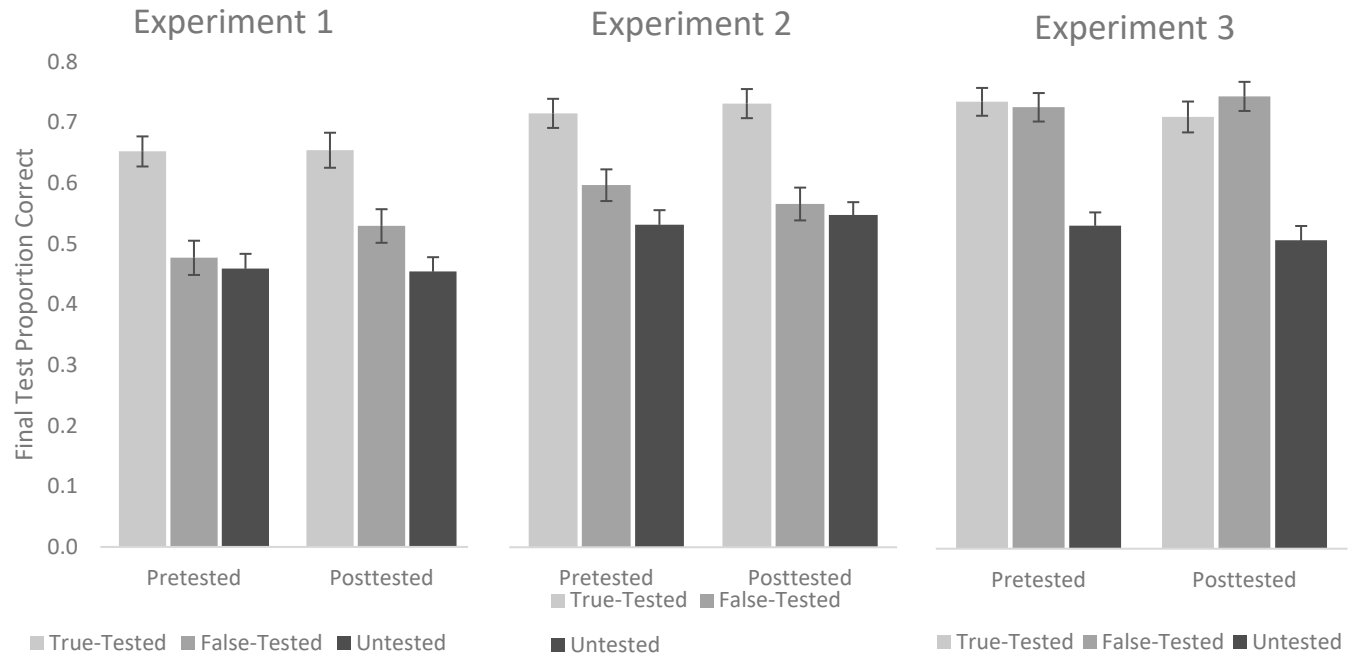


Figure 2. Final test performance across the three experiments as a function of Item Type and whether participants received a pretest or a posttest on a given passage. Error bars represent standard error of the mean.

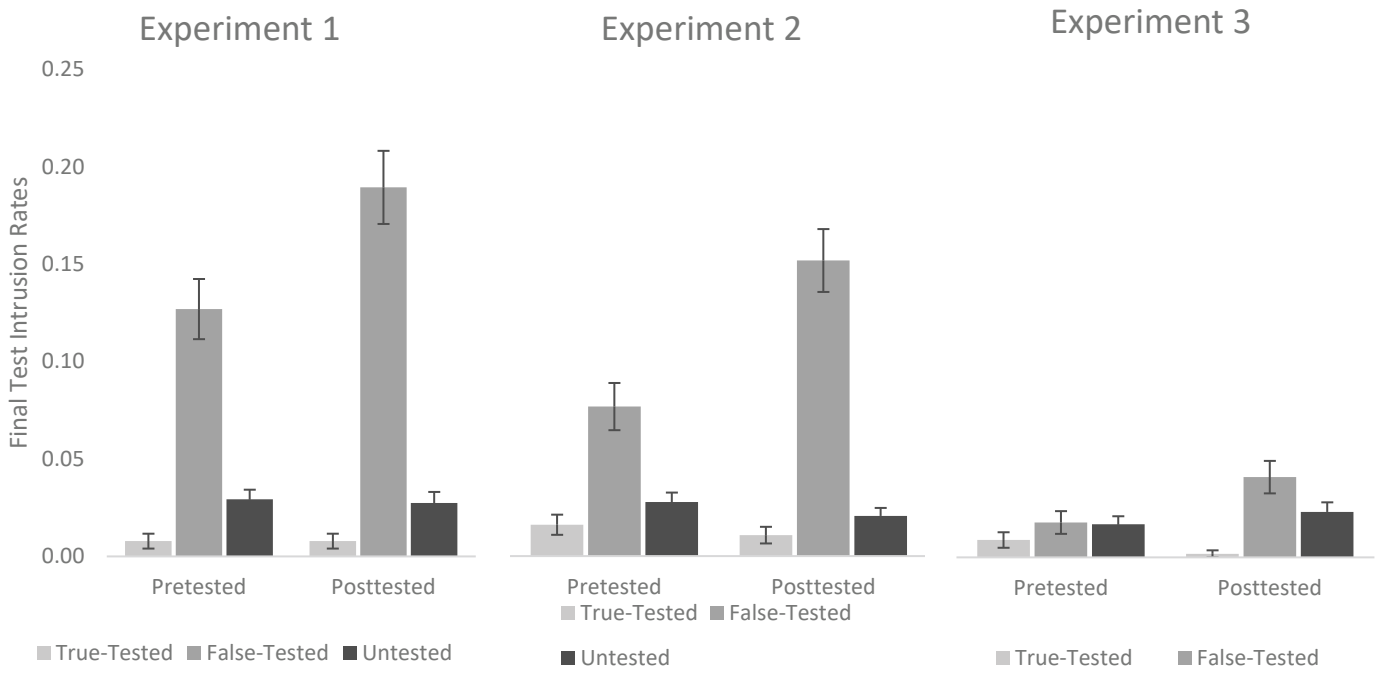


Figure 3. False intrusion rates across the three experiments as a function of Item Type and whether participants received a pretest or a posttest on a given passage. Error bars represent standard error of the mean.

Examining the simple effects of Item Type within each test condition revealed significant effects in both the Pretest condition, $F(2, 254) = 44.80$, $p < 0.001$, $\eta^2_p = 0.26$, and the Posttest condition, $F(2, 254) = 76.78$, $p < 0.001$, $\eta^2_p = 0.38$. Focusing first on the pretest condition, we found that the intrusion rate was significantly higher for False-Tested items ($M = 0.13$, $SD = 0.17$) than both Untested items ($M = 0.03$, $SD = 0.06$), $t(127) = 6.40$, $p < 0.001$, $d = 0.57$, 95% CI [0.38, 0.75], $BF_{01} < 0.001$, and True-Tested items ($M = 0.01$, $SD = 0.04$), $t(127) = 7.27$, $p < 0.001$, $d = 0.64$, 95% CI [0.45, 0.83], $BF_{01} < 0.001$, while being significantly

lower for True-Tested items than Untested items, $t(127) = 3.46$, $p < 0.001$, $d = 0.31$, 95% CI [0.13, 0.48], $BF_{01} = 0.050$. A similar pattern was observed in the Posttest condition: the intrusion rate was significantly higher for False-Tested items ($M = 0.19$, $SD = 0.21$) than both Untested items ($M = 0.03$, $SD = 0.06$), $t(127) = 8.45$, $p < 0.001$, $d = 0.75$, 95% CI [0.55, 0.94], $BF_{01} < 0.001$, and True-Tested items ($M = 0.01$, $SD = 0.04$), $t(127) = 9.56$, $p < 0.001$, $d = 0.85$, 95% CI [0.64, 1.05], $BF_{01} < 0.001$, while being significantly lower for True-Tested items than Untested items, $t(127) = 2.85$, $p = 0.005$, $d = 0.25$, 95% CI [0.08, 0.43], $BF_{01} = 0.287$.

3. Experiment 2

In Experiment 1, no differences were found on final test performance between items that were pretested versus posttested. Specifically, both types of tests appeared to promote recall performance at final test, and to an extent that did not significantly differ. However, a significantly higher proportion of intrusions did result from taking a posttest than taking a pretest, with this finding being observed despite posttest performance being significantly better than pretest performance. The goal of experiment 2 was to replicate experiment 1 and to examine how the addition of corrective feedback affects final test performance and intrusion rates.

Feedback has been demonstrated to be important for both pretesting and posttesting, though it is unclear whether it is equally important. By the nature of pretesting, the correct information is encountered after the pretest, so even without feedback, the final information encountered prior to the final test is the correct information. With posttesting, however, the order is study then test. As a consequence, for any "False" item, the last statement a participant encounters will be incorrect. Recent exposure to misinformation may therefore make posttesting especially prone to creating false memories—a disadvantage that corrective feedback may be able to help overcome. Such a finding would be consistent with prior work showing that the addition of feedback strengthens the posttesting benefit (Butler & Roediger, 2008; Kang et al., 2007). Indeed, McClusky (1934) suggested feedback as an antidote to the negative suggestion effect found in his experiments with True/False tests.

3.1. Methods

3.1.1. Participants and Design

A total of 140 UCSC undergraduates (M age: 19.86 years, range = 18–34; 110 female, 28 male, 1 write-in response, 1 declined to state) participated for partial credit in a psychology course as approved by the university's IRB. The study design was identical to that of Experiment 1. A power analysis, based on the effect size for the difference in intrusion rates between pretested and posttested false items in Experiment 1 ($d = 0.23$), suggested that 135 participants would be necessary for the experiment to have 80% power to observe a similar effect size in Experiment 2. As in Experiment 1, the study was initially conducted in person using OpenSesame ($n = 29$). To expedite data collection, however, we transitioned to run all subsequent participants remotely using the online survey platform, Qualtrics ($n = 111$). With this shift, we changed the distractor task from Sudoku to a word search.

3.1.2. Materials and Procedure

The materials and procedure for Experiment 2 were identical to that of Experiment 1, with the exception that corrective feedback was provided after each pre/posttest question. Specifically, after participants submitted their response, the statement "The correct answer is TRUE" or "The correct answer is FALSE" was displayed on screen until the participant pressed an arrow button to continue. This feedback was provided on every trial, regardless of the participant's accuracy in responding to the pre/posttest.

3.2. Results

3.2.1. Pretest and Posttest Performance

Similar to Experiment 1, we conducted a 2 (Test Type: Pretest vs. Posttest) by 2 (Question Type: True vs. False) repeated-measures ANOVA. As can be seen in the center panel of Figure 1, main effects were observed such that performance on the posttest (M = 0.73, SD = 0.19) was significantly higher than performance on the pretest (M = 0.56, SD = 0.20), $F(1, 139) = 55.13$, $p < 0.001$, $\eta^2_p = 0.28$., and performance on True Questions (M = 0.78, SD = 0.15) was significantly better than on False Questions (M = 0.52, SD = 0.20), $F(1, 139) = 218.40$, $p < 0.001$, $\eta^2_p = 0.61$. Moreover, a significant interaction was observed between Test Type and Question Type, $F(1, 139) = 50.68$, $p < 0.001$, $\eta^2_p = 0.27$. Follow-up tests revealed that although performance was better on True questions than False questions on both the pretest $t(139) = 15.12$, $p < 0.001$, $d = 1.28$, 95% CI [1.05–1.50], $BF_{01} < 0.001$, and posttest, $t(139) = 5.84$, $p < 0.001$, $d = 0.49$, 95% CI [0.32, 0.67], $BF_{01} < 0.001$, the difference was much greater on the pretest than on the posttest.

3.2.2. Final Test Performance

We conducted a 2 (Test Type: Pretest vs. Posttest) by 3 (Item Type: True-Tested vs. False-Tested vs. Untested) repeated-measures ANOVA on the final test performance. As can be seen in the center panel of Figure 2, a main effect of Test Type was not observed, $F(1, 139) = 0.00$, $p = 0.99$, $\eta^2_p = 0.00$. Specifically, final test performance did not differ depending on whether participants received a pretest (M = 0.61, SD = 0.24) or a posttest, (M = 0.61, SD = 0.23). A main effect of Item Type was observed, $F(2, 278) = 55.65$, $p < 0.001$, $\eta^2_p = 0.29$, however, such that True-Tested Items (M = 0.72, SD = 0.24) were recalled better than both False-Tested Items (M = 0.58, SD = 0.26), $t(139) = 6.93$, $p < 0.001$, $d = 0.59$, 95% CI [0.41, 0.76], $BF_{01} < 0.001$, and Untested items (M = 0.54, SD = 0.22), $t(139) = 10.55$, $p < 0.001$, $d = 0.89$, 95% CI [0.69, 1.09], $BF_{01} < 0.001$. False-Tested items were also recalled better than Untested items, $t(139) = 2.51$, $p = 0.013$, $d = 0.21$, 95% CI [0.04, 0.38], $BF_{01} = 0.880$. Finally, we failed to detect a significant interaction, $F(2, 278) = 1.19$, $p = 0.31$, $\eta^2_p = 0.01$, and thus again failed to find evidence that performance on the pretest and posttest differed as a function of Item Type.

3.2.3. Final Test Intrusions

We conducted a 2 (Test Type: Pretested vs. Posttested) by 3 (Item Type: True-Tested vs. False-Tested vs. Untested) repeated-measures ANOVA on the proportion of intrusions on the final test. As in Experiment 1, this analysis revealed significant main effects of both Item Type, $F(2, 278) = 67.32$, $p < 0.001$, $\eta^2_p = 0.33$, and Test Type, $F(1, 139) = 10.29$, $p = 0.002$, $\eta^2_p = 0.07$, which were in turn qualified by a significant interaction, $F(2, 278) = 15.10$, $p < 0.001$, $\eta^2_p = 0.10$. As can be seen in the center panel of Figure 3, when we compare the critical False-Tested items directly across the two test types, participants produced significantly more intrusions on items that were false on the posttest (M = 0.15, SD = 0.19) than they did for items that were false on the pretest (M = 0.08, SD = 0.06), $t(139) = 4.04$, $p < 0.001$, $d = 0.34$, 95% CI [0.17, 0.51], $BF_{01} = 0.007$.

Examining the simple effects of Item Type separately for each test condition revealed significant differences in both the Pretest condition, $F(2, 278) = 15.97$, $p < 0.001$, $\eta^2_p = 0.10$, and the Posttest condition, $F(2, 278) = 62.61$, $p < 0.001$, $\eta^2_p = 0.31$. Focusing first on the pretest condition, we found that the intrusion rate was significantly higher for False-Tested items (M = 0.08, SD = 0.06) than both Untested items (M = 0.03, SD = 0.06), $t(139) = 3.70$, $p < 0.001$, $d = 0.31$, 95% CI [0.14, 0.48], $BF_{01} = 0.023$, and True-Tested items (M = 0.02, SD = 0.06), $t(139) = 4.71$, $p < 0.001$, $d = 0.40$, 95% CI [0.23, 0.57], $BF_{01} = 0.001$. Unlike in Experiment 1, a significant difference was not observed between True-Tested items and

Untested items, $t(139) = 1.68$, $p = 0.096$, $d = 0.14$, 95% CI [−0.03, 0.31], $BF_{01} = 3.758$. In the Posttest condition, the intrusion rate was significantly higher for False-Tested items ($M = 0.15$, $SD = 0.19$) than both Untested items ($M = 0.02$, $SD = 0.05$), $t(139) = 7.82$, $p < 0.001$, $d = 0.66$, 95% CI [0.48, 0.84], $BF_{01} < 0.001$, and True-Tested items ($M = 0.01$, $SD = 0.05$), $t(139) = 8.37$, $p < 0.001$ $d = 0.71$, 95% CI [0.52, 0.89], $BF_{01} < 0.001$, with once again a significant difference not being observed between True-Tested items and Untested items, items $t(139) = 1.82$, $p = 0.07$, $d = 0.15$, 95% CI [−0.01, 0.32], $BF_{01} = 2.921$.

4. Experiment 3

Overall, the patterns of results observed in Experiment 2 were largely the same as those observed in Experiment 1. Despite performance on the posttest being better than performance on the pretest, both tests led to comparable learning benefits relative to untested items, and both led to false intrusions. Most critically, posttesting led to more false intrusions than pretesting, with the effect size observed in Experiment 2 being just as large as it was in Experiment 1. Although corrective feedback did lead to lower intrusion rates overall (which can be seen visually in Figure 3), it did not appear to do so in a way that was more effective at reducing false intrusions following posttesting than after pretesting.

In Experiment 3 we examined whether more substantive feedback would be more effective in reducing the number of false intrusions resulting from posttests. Some prior research on errorful learning (e.g., Pashler et al., 2005) has demonstrated that more in-depth feedback, such as supplying answers rather than simple correct/incorrect feedback, may be critical for improving learning from errors. Whereas corrective feedback in Experiment 2 may have been effective at focusing attention towards the correct to-be-learned information following a pretest, it may not have been as effective following a posttest since there was not any opportunity to go back and study the to-be-learned information. When given substantive feedback, however, this difference may be mitigated. Providing learners with the correct to-be-learned information following a posttest may help reduce the likelihood that they will go on to intrude false information at final recall. Indeed, it may do so to such an extent that posttesting will no longer lead to rates of false intrusion greater than that observed following pretesting.

4.1. Methods

4.1.1. Participants and Design

A total of 140 UCSC undergraduates (*M* age: 20.57 years, range = 18–50; 106 female, 27 male, 6 write-in responses, 1 declined to state) participated for partial credit in a psychology course as approved by the university's IRB. The study design was identical to that of the first two experiments. Considering the similar structure to the previous experiments, we reused the power analysis from Experiment 2 with a goal of 135 participants. The experiment was run entirely online using Qualtrics.

4.1.2. Materials and Procedure

The materials and procedure were identical to that of Experiment 2, with the exception that an additional sentence explaining the correct information was added to the feedback for each pre/posttest question. For example, after responding to a pre/posttest question such as "As part of her mission, Joan of Arc took a vow of silence." the participant would receive the feedback: "The correct answer is FALSE; Joan took a vow of *chastity*." For true items, the feedback similarly restated the correct fact. This additional sentence was added to the feedback regardless of whether the answer was true or false and regardless of whether the participant responded correctly. See the rightmost column of Appendix A for a full list of the feedback statements for each item.

4.2. Results

4.2.1. Pretest and Posttest Performance

As in the prior experiments, we began by conducting a 2 (Test Type: Pretest vs. Posttest) by 2 (Question Type: True vs. False) repeated-measures ANOVA. As can be seen in the right panel of Figure 1, main effects were observed such that performance on the posttest (M = 0.74, SD = 0.19) was significantly better than on the pretest (M = 0.57, SD = 0.18), $F(1, 139) = 59.72$, $p < 0.001$, $\eta^2_p = 0.30$, and performance on the True Questions (M = 0.79, SD = 0.15) was significantly better than performance on the False Questions (M = 0.52, SD = 0.20), $F(1, 139) = 202.91$, $p < 0.001$, $\eta^2_p = 0.59$. A significant interaction was also found between Test Type and Question Type, $F(1, 139) = 31.55$, $p < 0.001$, $\eta^2_p = 0.19$. Follow up tests revealed once again that although performance was greater on True questions than False questions on both the pretest $t(139) = 14.55$, $p < 0.001$, $d = 1.23$, 95% CI [1.01–1.45], $BF_{01} < 0.001$, and the posttest, $t(139) = 6.35$, $p < 0.001$, $d = 0.54$, 95% CI [0.36, 0.71], $BF_{01} < 0.001$, the difference was much greater on the pretest than on the posttest.

4.2.2. Final Test Performance

We conducted a 2 (Test Type: Pretest vs. Posttest) by 3 (Item Type: True-Tested vs. False-Tested vs. Untested) repeated-measures ANOVA on final test performance. As in the prior experiments, a main effect of Test Type was not observed, $F(1, 139) = 0.35$, $p = 0.55$, $\eta^2_p = 0.00$, with final test performance failing to differ as a function of whether participants received a pretest (M = 0.67, SD = 0.21) or a posttest (M = 0.66, SD = 0.23). A main effect of Item Type was observed, $F(2, 278) = 81.01$, $p < 0.001$, $\eta^2_p = 0.37$. Interestingly, True-Tested items (M = 0.72, SD = 0.25) were recalled better than Untested items (M = 0.52, SD = 0.21), $t(139) = 10.43$, $p < 0.001$, $d = 0.88$, 95% CI [0.69, 1.08], $BF_{01} < 0.001$, but they were not recalled better than False-Tested items (M = 0.74, SD = 0.23), $t(139) = 0.64$, $p = 0.520$, $d = 0.05$, 95% CI [−0.11, 0.22], $BF_{01} = 12.154$. False-Tested items were recalled better than Untested items, $t(139) = 11.82$, $p < 0.001$, $d = 1.00$, 95% CI [0.79, 1.20], $BF_{01} < 0.001$. Finally, we once again failed to observe a significant interaction, $F(2, 278) = 1.08$, $p = 0.34$, $\eta^2_p = 0.01$, and thus again failed to find evidence that performance on the pretest and posttest differed as a function of Item Type. Final test performance for Experiment 3 can be seen in the right panel of Figure 2.

4.2.3. Final Test Intrusions

As in prior experiments, we conducted a 2 (Test Type: Pretested vs. Posttested) by 3 (Item Type: True-Tested vs. False-Tested vs. Untested) repeated-measures ANOVA on the proportion of intrusions on the final test. This analysis revealed a significant main effect of Item Type, $F(2, 278) = 12.32$, $p < 0.001$, $\eta^2_p = 0.08$, but not of Test Type, $F(1, 139) = 3.25$, $p = 0.074$, $\eta^2_p = 0.02$. Most importantly, a significant interaction was found between Test Type and Item Type, $F(2, 278) = 4.15$, $p = 0.017$, $\eta^2_p = 0.03$. As in the previous experiments, we began by comparing the critical False-Tested items directly across the two test types. Once again, participants produced significantly more intrusions on items that were false on the Posttest (M = 0.04, SD = 0.10) than they did for items that were false on the Pretest (M = 0.02, SD = 0.07), $t(139) = 2.23$, $p = 0.027$, $d = 0.19$, 95% CI [0.02, 0.36], $BF_{01} = 1.324$. As can be seen in the right panel of Figure 3, however, overall intrusion rates in Experiment 3 were substantially lower than in the previous experiments. It is worth noting that despite reaching the threshold for statistical significance using null-hypothesis testing, the Bayes Factor actually suggested nominally more evidence for the null. As a follow-up analysis, we combined data from all three experiments, and with this more highly powered comparison, we found strong evidence that participants produced significantly more intrusions on items that were false on the Posttest (M = 0.13, SD = 0.18) than they did for items that were false

on the Pretest (M = 0.07, SD = 0.07), $t(407) = 5.19$, $p < 0.001$, $d = 0.26$, 95% CI [0.16, 0.36], $BF_{01} < 0.001$.

Examining the simple effects of Item Type separately for each test condition revealed significant differences in the posttest condition, $F(2, 278) = 12.81$, $p < 0.001$, $\eta^2_p = 0.08$, but not in the pretest condition, $F(2, 278) = 0.98$, $p = 0.378$, $\eta^2_p = 0.01$. This pattern suggests that the interaction was driven by substantive feedback being successful at eliminating the negative suggestion effective in the pretest condition, but not in the posttest condition. Indeed, the substantive feedback condition was so effective that it completely eliminated any significant differences in intrusion rates; intrusions for False-Tested items (M = 0.02, SD = 0.07) were reduced to the same floor-level as True-Tested items (M = 0.01, SD = 0.05) and Untested items (M = 0.02, SD = 0.04). Although dramatically reduced as well, small but significant differences remained in the posttest condition. Specifically, intrusion rates were still higher for False-Tested items (M = 0.04, SD = 0.10) than they were for either True-Tested items (M = 0.00, SD = 0.02), $t(139) = 4.62$, $p < 0.001$, $d = 0.39$, 95% CI [0.22, 0.56], $BF_{01} = 0.001$, or Untested items (M = 0.02, SD = 0.05), $t(139) = 2.23$, $p = 0.027$, $d = 0.19$, 95% CI [0.02, 0.36], $BF_{01} = 1.307$. Significantly fewer intrusions were observed on the True-Tested items than on the Untested items, $t(139) = 3.83$, $p < 0.001$, $d = 0.32$, 95% CI [0.15, 0.49], $BF_{01} = 0.015$.

5. General Discussion

Pretesting and posttesting have both been shown to be effective methods of improving final test performance (Karpicke & Roediger, 2008; Richland et al., 2009). Although many studies have focused on the benefits of either pretesting or posttesting alone, there have been relatively few comparing the two directly under the same set of conditions (Geller et al., 2017; McDaniel et al., 2011; Latimier et al., 2019; Pan & Sana, 2021). The present experiments add to this body of research by comparing pretesting and posttesting through the lens of True/False tests. In addition to examining the relative effectiveness of pretesting or posttesting for promoting subsequent recall performance, we also examined intrusion rates. Intrinsic to the nature of True/False tests is that test-takers will be exposed to false information. We sought to examine the extent to which such exposure leads to a negative suggestion effect (e.g., Remmers & Remmers, 1926) and whether this effect would be larger in the case of pretesting or posttesting.

In three experiments, participants read two educational passages, with one passage being preceded by a pretest, and the other being followed by a posttest. The pretest and posttest each consisted of a True/False test in which some items were true and other items were false. Not surprisingly, participants performed better on the posttest than they did on the pretest, a finding which makes sense given that participants had the opportunity to study the passage before taking the posttest. Nevertheless, when participants were given a final cued recall test, performance failed to differ as a function of test type, with both significantly enhancing performance relative to untested items. A significant difference was observed, however, in terms of intrusions. Specifically, posttesting led to significantly more intrusions than pretesting, with participants recalling the false information they were exposed to in the posttest significantly more often than the false information they were exposed to in the pretest. This difference was observed when participants were not given any feedback on the True/False tests (Experiment 1), when participants were given corrective feedback (Experiment 2), and when participants were given substantive feedback (Experiment 3). In fact, although substantive feedback effectively eliminated the negative suggestion following pretesting, it did not do so following posttesting.

A key finding from our experiments was that final test performance failed to differ as a function of pretesting and posttesting. This result is particularly interesting as it stands in contrast to other investigations, which have found either a pretesting advantage

(e.g., Pan & Sana, 2021) or a posttesting advantage (e.g., McDaniel et al., 2011). One interpretation is that the relative efficacy of pretesting and posttesting is not absolute but rather depends on specific methodological factors, such as the format of the test questions, the extent to which pretesting enhances subsequent study behaviors, and the relative success of retrieval practice in posttesting. A pretesting advantage, for example, might emerge when test formats prompt deeper or more elaborative encoding during study, whereas a posttesting advantage may appear when conditions favor the successful retrieval practice and consolidation of recently learned material. Given the number of potential moderators, we are hesitant to overinterpret our null results. What our findings do suggest is that at least under the conditions we employed, taking a True/False test has the power to enhance learning, and it can do so regardless of whether it is administered before or after studying a to-be-learned passage.

It is noteworthy that learning from initially false statements was significantly less effective than learning from initially true statements. Across Experiments 1 and 2, while testing with true items provided a robust benefit to final performance, the benefit from testing with false items was varied and more limited. This pattern of results presents somewhat of a pedagogical challenge. Presumably, if a test consisted of only true items, the learning benefits from such items would be diminished, since learners would not have to think carefully about whether a given statement was true or false. Critically, in Experiment 3, the false items were found to benefit substantially from True/False testing, and to an extent that did not differ from that of true items. This finding suggests that all items have the potential to benefit from True/False testing as long as substantive feedback is provided.

Looking beyond overall test performance, our results demonstrate that taking a True/False test can put learners at risk for a negative suggestion effect. Specifically, exposure to incorrect information on a True/False test can make participants more likely to recall that information on the final test. This finding extends prior research demonstrating that exposure to incorrect lures on multiple-choice tests can increase their subsequent endorsement (e.g., Roediger & Marsh, 2005). Moreover, in the case of posttesting, a negative suggestion effect was observed even when participants were given substantive feedback following the True/False test. Specifically, participants continued to output the false information even when they were told that the information was false and given the correct information prior to taking the final test, albeit at a relatively low rate. The more important observation for present purposes was that across all three experiments, the negative suggestion effect was significantly larger following a True/False posttest than following a True/False pretest. Participants were consistently more likely to intrude incorrect information learned from exposure to false items on a posttest than they were from false items on a pretest, with this difference emerging even though participants were less likely to select incorrect information on the posttests than the pretests, and even though (as discussed above) posttesting led to overall final test performance that was not significantly different than pretesting. This pattern of results raises a critical theoretical question: what cognitive mechanisms make posttesting particularly vulnerable to the negative suggestion effect?

One plausible explanation centers on the recency of the misinformation. In Experiment 1, for example, because no feedback was given, the false statement on a posttest was the last piece of related information a participant encountered, which could have made it more accessible at the final test. This simple recency account, however, is challenged by the results of Experiment 3. In that experiment, substantive feedback presented the correct information after the false statement, yet the posttest intrusion rate remained higher than the pretest intrusion rate. Although recency may play a role, it seems unlikely to be the sole explanation.

A more convincing explanation may lie in the kind of cognitive mindset fostered by the two procedures. When taking a pretest, participants may undertake the pretest with a more evaluative or exploratory mindset. In this state, they may process the True/False questions critically, searching for clues and treating the information with a degree of skepticism that inoculates them against encoding the incorrect information in a way that would make that information likely to be retrieved as correct at final test. In contrast, because posttesting occurs after study, a learner's goal is to try to retrieve accurate information and to do so in a way that is integrated with other information and knowledge about the topic. This could lead to higher intrusion rates by making false information more likely to be integrated into long-term memory.

Moreover, from a source monitoring perspective, participants may be better able to take advantage of context cues following pretesting than posttesting. In the case of pretesting, even if participants remember selecting a statement as being True, they can infer that because they made that selection prior to studying the to-be-learned passage, they should approach the veracity of that statement with some caution. In the case of posttesting, however, participants do not have this same benefit. If they remember selecting a given statement as True, they cannot use the general placement of the True/False test as an indication of the extent to which they were making that selection based on a guess or their memory for the information they had just studied. Indeed, posttesting may generally lead to more source confusion than pretesting, especially if participants are activating information from the passage while taking the posttest.

Finally, posttesting may lead to more intrusions than pretesting owing to dynamics related to retrieval-induced forgetting. Research has shown that retrieving some information can cause the forgetting of other information (Anderson et al., 1994; Bäuml & Kliegl, 2017; Storm & Levy, 2012). In the current context, a posttest TF false item may serve as a retrieval practice event, prompting the forgetting of the related true information. Consequently, when given the cued recall final test, participants may have been less likely to recall the true information, and therefore more likely to output the false information for those questions than they would have been otherwise. One might note that final recall performance of true information was not impaired by exposure to false TF items, but this observation does not negate the possibility that at least some of the true information experienced retrieval-induced forgetting. It is possible, for example, that whereas some true information benefited from posttesting via retrieval-induced facilitation (Bäuml & Samenieh, 2012; Chan et al., 2006; Oliva & Storm, 2023), other true information was impaired, and it was this latter case that contributed to the increase in false intrusions. Unfortunately, given the way the study was designed, it is impossible to separate the effects of posttesting on memory for false information from that of true information, but this is something future research may consider addressing more directly.

From a more applied perspective, the current results suggest that educators have good reason to be concerned about the negative suggestion effect in the context of True/False tests, especially when learners are not given feedback. In Experiment 1, participants recalled the False information 19% of the time, an intrusion rate that was lowered to 13% and 4% when corrective and substantive feedback were given, respectively. In contrast, pretesting led to significantly lower intrusion rates in all three studies (13%, 8%, and 2%, respectively) even though it was just as effective in yielding equivalent levels of overall final test performance. As such, our results suggest that if educators want to implement True/False practice tests to enhance learning, but are concerned about negative suggestion effects, pretesting may be the safer pedagogical choice. It provides a comparable benefit with a significantly lower risk of instilling false information. In fact, we failed to find any evidence of a negative suggestion effect in the pretesting condition of Experiment 3. When

participants were given substantive feedback, they were no more likely to recall the false information on the final test than they would have been had they not been exposed to that false information at all. Of course, it is possible that following a more extended delay, a negative suggestion effect would re-emerge, similar to what has been observed in previous work on the sleeper effect in memory (Jacoby et al., 1989).

A few important limitations are worth noting, as they provide important context for our findings and suggest potential directions for future research. First, our participants consisted of undergraduate students at a university, and it remains to be seen whether the same patterns of intrusions would be observed in other populations, such as K-12 students in classroom settings. Second, our materials were limited to two short educational passages. The mechanisms may differ with more complex materials or as a function of prior knowledge or how related the false information is to the true information. Third, we encourage future research to examine the consequences of True/False testing following longer retention intervals, as the effects observed after a short delay may not necessarily be observed after a longer delay. Finally, True/False are typically not used in isolation. Rather, they are embedded within tests consisting of other kinds of items (e.g., multiple choice), and how participants respond and learn from True/False tests may differ as a function of such context.

In conclusion, the present research adds a new dimension to the discussion of the relative benefits of pretesting and posttesting. Although both pretesting and posttesting have the power to enhance final test performance (observed here in the context of True/False test-taking), posttesting appears to be most vulnerable to the negative suggestion effect. Investigating more precisely when and why this is the case represents a critical area for future theoretical and applied research in the science of learning.

Author Contributions: Conceptualization, K.K.J.; methodology, K.K.J.; formal analysis, K.K.J.; investigation, K.K.J.; data curation, K.K.J.; writing—original draft preparation, K.K.J.; writing—review and editing, K.K.J. and B.C.S.; visualization, K.K.J.; supervision, B.C.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was approved by the Institutional Review Board of University of California, Santa Cruz, HS1919, exempt (original approval: 27 September 2012) and HS3756, exempt (original approval: 26 October 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Passage: Joan of Arc				
	True Item	**False Item**	**Cued Recall Question**	**Substantive Feedback (Exp 3 Only)**
1	Before Joan of Arc convinced the embattled crown prince Charles of Valois to allow her to lead an army, she had no military experience.	Before Joan of Arc convinced the embattled crown prince Charles of Valois to allow her to lead an army, she had 5 years of military experience.	Before Joan of Arc convinced the embattled crown prince Charles of Valois to allow her to lead an army, how much military experience did she have? _____.	She had *no military experience*.

Passage: Joan of Arc				
	True Item	**False Item**	**Cued Recall Question**	**Substantive Feedback (Exp 3 Only)**
2	In 1920, Joan of Arc was officially canonized as a saint.	In 1920, Joan of Arc was officially recognized on a list of famous female warriors.	In 1920, Joan of Arc was officially _____.	She was *canonized*.
3	Joan of Arc was born around the year 1412.	Joan of Arc was born around the year 1300	Around what year was Joan of Arc born? _____.	She was born around *1412*.
4	Joan of Arc's father worked as a farmer.	Joan of Arc's father worked as a tailor.	Her father worked as a _____.	He worked as a *farmer*.
5	Joan's mother instilled in her a deep love for the Catholic Church.	Joan's mother instilled in her a deep love for books and scholarly pursuits.	Joan's mother instilled in her a deep love for _____.	She instilled a love for *the church*.
6	In 1420, when the French crown prince was disinherited, King Henry V was made ruler of England and France.	In 1420, when the French crown prince was disinherited, King Edward was made ruler of England and France.	In 1420, when the French crown prince was disinherited, _____ was made ruler of England and France.	King *Henry V* was made ruler.
7	Joan heard voices, which she believed were coming from God.	Joan heard voices, which she believed were coming from her deceased uncle.	Where did Joan believe the voices she heard were coming from?	She believed they were coming from *God*.
8	Joan went to court after her father tried to arrange a marriage for her.	Joan went to court after her father began to physically abuse her.	Joan went to court after her father _____.	He tried to *arrange a marriage* for her.
9	As part of her mission, Joan took a vow of chastity.	As part of her mission, Joan took a vow of silence.	As part of her mission, Joan took a vow of _____.	She took a vow of *chastity*.
10	Joan dressed in men's clothes to make the 11-day journey across enemy territory to Chinon, the site of the crown prince's palace.	Joan dressed in a black dress to make the 11-day journey across enemy territory to Chinon, site of the crown prince's palace.	Joan dressed in _____ to make the 11-day journey across enemy territory to Chinon, site of the crown prince's palace.	She dressed in *men's clothes*.
11	Joan promised Charles she would see him crowned king at Reims, the traditional site of French royal investiture.	Joan promised Charles she would see him crowned king at Versailles, the traditional site of French royal investiture.	Joan promised Charles she would see him crowned king at _____, the traditional site of French royal investiture.	Charles would be crowned at *Reims*.
12	After Joan's early victory, she and her followers escorted Charles across enemy territory, enabling him to be coronated in July 1429.	After Joan's early victory, she and her followers escorted Charles across enemy territory, enabling him to be freed from enemy capture in July 1429.	After Joan's early victory, she and her followers escorted Charles across enemy territory, enabling him to be _____ in July 1429.	Charles was *coronated* at that time.
13	In the spring of 1430, the king ordered Joan to confront an assault on Compiégne carried out by the Anglo-Burgundians.	In the spring of 1430, the king ordered Joan to confront an assault on Compiégne carried out by the Austrians.	In the spring of 1430, the king ordered Joan to confront an assault on Compiégne carried out by the _____.	The assault was carried out by the *Anglo-Burgundians*.

Passage: Joan of Arc				
	True Item	**False Item**	**Cued Recall Question**	**Substantive Feedback (Exp 3 Only)**
14	When Joan was captured, she was forced to answer some 70 charges against her, including witchcraft, heresy, and dressing like a man.	When Joan was captured, she was forced to answer some 15 charges against her, including witchcraft, heresy, and dressing like a man.	When Joan was captured, she was forced to answer some _____ (#) charges against her, including witchcraft, heresy, and dressing like a man.	She had to answer *70* charges.
15	After a year in captivity and under threat of death, Joan relented and signed a confession saying she had never received divine guidance.	After a year in captivity and under threat of death, Joan relented and signed a confession saying she had captured and killed soldiers.	After a year in captivity and under threat of death, Joan relented and signed a confession saying _____.	She signed a confession saying that she *did not receive divine guidance*.
16	Joan of Arc was 19 years old when she was burned at the stake.	Joan of Arc was 25 years old when she was burned at the stake.	How old was Joan when she was burned at the stake? _____	She was *nineteen*.

Passage: Machu Picchu				
	True Item	**False Item**	**Cued Recall Question**	**Substantive Feedback (Exp 3 Only)**
1	Machu Picchu is tucked away in the rocky countryside northwest of Cuzco, Peru.	Machu Picchu is tucked away in the rocky countryside northwest of Nazca, Peru.	Machu Picchu is tucked away in the rocky countryside northwest of _____, Peru.	It is near *Cuzco*, *Peru*.
2	In the 16th century, the civilization that created Machu Picchu was nearly wiped out by Spanish Invaders.	In the 16th century, the civilization that created Machu Picchu was nearly wiped out by famine.	In the 16th century, the civilization that created Machu Picchu was nearly wiped out by _____.	It was nearly wiped out by *Spanish Invaders*.
3	Hiram Bingham is credited with discovering Machu Picchu in 1911.	Howard Carter is credited with discovering Machu Picchu in 1911.	_____ is credited with discovering Machu Picchu in 1911.	*Hiram Bingham* is credited with this discovery.
4	The site stretches over an impressive 5-mile distance, featuring more than 3000 stone steps.	The site stretches over an impressive 5-mile distance, featuring more than 3000 buildings.	The site stretches over an impressive 5-mile distance, featuring more than 3000 _____.	It features more than 3000 *stone steps*.
5	It was abandoned an estimated 100 years after its construction.	It was abandoned an estimated 1000 years after its construction.	It was abandoned an estimated _____ years after its construction.	It was abandoned after *100 years*.
6	One theory of why Machu Picchu was deserted was that there was a smallpox epidemic.	One theory of why Machu Picchu was deserted was that there was a Malaria epidemic.	One theory of why Machu Picchu was deserted was that there was a/n _____ epidemic.	The theory states that there was a *smallpox epidemic*.

Passage: Machu Picchu				
	True Item	**False Item**	**Cued Recall Question**	**Substantive Feedback (Exp 3 Only)**
7	When the archaeologist credited with finding Machu Picchu arrived, he was actually trying to find Vilcabamba, the last Inca stronghold.	When the archaeologist credited with finding Machu Picchu arrived, he was actually trying to find Saqsaywaman, the first capital of Peru.	When the archaeologist credited with finding Machu Picchu arrived, he was actually trying to find _____.	He was trying to find *Vilcabamba*, *the last Inca stronghold*.
8	The name Machu Picchu means Old Peak in the native Quechua language.	The name Machu Picchu means Green Tree in the native Quechua language.	The name Machu Picchu means _____ in the native Quechua language.	The name means *Old Peak*.
9	The person who helped the archaeologist find Machu Picchu was 11 years old.	The person who helped the archaeologist find Machu Picchu was 80 years old.	How old was the person who helped the archaeologist find the site?	He was *11*.
10	When the excavated artifacts were brought to Yale University, a custody dispute was ignited.	When the excavated artifacts were brought to Columbia University, a custody dispute was ignited.	When the excavated artifacts were brought to _____ University, a custody dispute was ignited.	The artifacts were brought to *Yale*.
11	The artifacts of Machu Picchu were finally returned to Peru under Barack Obama.	The artifacts of Machu Picchu were finally returned to Peru under Bill Clinton.	Under which president were the artifacts finally returned to Peru?	They were returned under *Barack Obama*.
12	In the midst of a tropical mountain forest on the eastern slopes of the Andes mountains, Machu Picchu blends seamlessly into its natural setting.	In the midst of a tropical mountain forest on the eastern slopes of the Cordillera mountains, Machu Picchu blends seamlessly into its natural setting.	In the midst of a tropical mountain forest on the eastern slopes of the _____, Machu Picchu blends seamlessly into its natural setting.	Machu Picchu is along the slopes of the *Andes* Mountains.
13	Its central buildings are prime examples of a masonry technique mastered by the Incas in which stones fit together without mortar.	Its central buildings are prime examples of a masonry technique mastered by the Incas in which stones were cut in interesting shapes.	Its central buildings are prime examples of a masonry technique mastered by the Incas in which stones _____.	The stones were cut *to fit without mortar*.
14	The Intihuatana stone, is a sculpted granite rock that is believed to have functioned as a solar clock.	The Intihuatana stone, is a sculpted granite rock that is believed to have functioned as an idol for worshipping a god.	The Intihuatana stone, is a sculpted granite rock that is believed to have functioned as a _____.	It is believed to have functioned as a *solar clock or calendar*.
15	In 2007, Machu Picchu was designated as one of the new 7 wonders of the world.	In 2007, Machu Picchu was designated as one of the new sacred lands of Peru.	In 2007, Machu Picchu was designated as one of the new _____.	It was designated as one of the new 7 world wonders.
16	Machu Picchu is the most famous ruin in South America, welcoming hundreds of thousands of people per year.	Machu Picchu is the most famous ruin in South America, welcoming around ten thousand people per year.	Machu Picchu is the most famous ruin in South America, welcoming _____ (#) people per year.	The site welcomes *hundreds of thousands* of people per year

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1063–1087. [CrossRef]

Bäuml, K.-H. T., & Kliegl, O. (2017). Retrieval-induced remembering and forgetting. In J. T. Wixted, & J. H. Byrne (Eds.), *Cognitive psychology of memory, Vol. 2 of learning and memory: A comprehensive reference* (pp. 27–51). Academic Press.

Bäuml, K.-H. T., & Samenieh, A. (2012). Selective memory retrieval can impair and improve retrieval of other memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 488–494. [CrossRef]

Brabec, J. A., Pan, S. C., Bjork, E. L., & Bjork, R. A. (2021). True-false testing on trial: Guilty as charged or falsely accused? *Educational Psychology Review*, *33*, 667–692. [CrossRef]

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616. [CrossRef]

Carpenter, S. K., & Toftness, A. R. (2017). The effect of prequestions on learning from video presentations. *Journal of Applied Research in Memory and Cognition*, *6*(1), 104–109. [CrossRef]

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 431–437. [CrossRef]

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*(4), 553–571. [CrossRef]

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology*, *21*(6), 919–940. [CrossRef]

Geller, J., Carpenter, S. K., Lamm, M. H., Rahman, S., Armstrong, P. I., & Coffman, C. R. (2017). Prequestions do not enhance the benefits of retrieval in a STEM classroom. *Cognitive Research: Principles and Implications*, *2*(1), 1–13. [CrossRef]

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*(4), 505–513. [CrossRef] [PubMed]

Hildenbrand, L., & Wiley, J. (2025). Supporting comprehension: The advantages of multiple–choice over true—False practice tests. *Memory & Cognition*, 1–17. [CrossRef]

Jacoby, L. L., Kelley, C., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology*, *56*(3), 326–338. [CrossRef]

James, K. K., & Storm, B. C. (2019). Beyond the pretesting effect: What happens to the information that is not pretested? *Journal of Experimental Psychology: Applied*, *25*(4), 576–587. [CrossRef]

Kang, S. H., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. [CrossRef]

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968. [CrossRef]

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 106–114. [CrossRef]

Kriechbaum, V. M., & Bäuml, K.-H. T. (2024). Retrieval practice can promote new learning with both related and unrelated prose materials. *Journal of Applied Research in Memory and Cognition*, *13*, 319–328. [CrossRef]

Latimier, A., Riegert, A., Peyre, H., Ly, S. T., Casati, R., & Ramus, F. (2019). Does pre-testing promote better retention than post-testing? *NPJ Science of Learning*, *4*(1), 1–15. [CrossRef] [PubMed]

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communications Monographs*, *66*(2), 125–144. [CrossRef]

Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, *43*(1), 14–26.

Little, J. L., & Bjork, E. L. (2016). Multiple-choice pretesting potentiates learning of related information. *Memory & Cognition*, *44*(7), 1085–1101. [CrossRef]

McClusky, H. Y. (1934). The negative suggestion effect of the false statement in the true-false test. *The Journal of Experimental Education*, *2*(3), 269–273. [CrossRef]

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414. [CrossRef]

Oliva, M. T., & Storm, B. C. (2023). Examining the effect size and duration of retrieval-induced facilitation. *Psychological Research*, *87*, 1166–1179. [CrossRef] [PubMed]

Overoye, A. L., James, K. K., & Storm, B. C. (2021). A little can go a long way: Giving learners some context can enhance the benefits of pretesting. *Memory*, *29*(9), 1206–1215. [CrossRef] [PubMed]

Pan, S. C., & Sana, F. (2021). Pretesting versus posttesting: Comparing the pedagogical benefits of errorful generation and retrieval practice. *Journal of Experimental Psychology. Applied*, *27*(2), 237–257. [CrossRef] [PubMed]

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 3–8. [CrossRef]

Pastötter, B., & Bäuml, K.-H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology*, *5*, 83305. [CrossRef]

Remmers, H. H., & Remmers, E. M. (1926). The negative suggestion effect on true-false examination questions. *Journal of Educational Psychology*, *17*(1), 52–56. [CrossRef]

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243. [CrossRef]

Roediger, H. L., III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011a). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382. [CrossRef]

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. [CrossRef]

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1155–1159. [CrossRef]

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011b). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation*, *55*, 1–36.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. [CrossRef]

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*(9), 641–656. [CrossRef]

Storm, B. C., & Levy, B. J. (2012). A progress report on the inhibitory account of retrieval-induced forgetting. *Memory & Cognition*, *40*, 827–843. [CrossRef] [PubMed]

Toppino, T. C., & Brochin, H. A. (1989). Learning from tests: The case of true-false examinations. *The Journal of Educational Research*, *83*(2), 119–124. [CrossRef]

Toppino, T. C., & Luipersbeck, S. M. (1993). Generality of the negative suggestion effect in objective tests. *The Journal of Educational Research*, *86*(6), 357–362. [CrossRef]

Uner, O., Tekin, E., & Roediger, H. L. (2022). True-false tests enhance retention relative to rereading. *Journal of Experimental Psychology: Applied*, *28*(1), 114–129. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Interpolated Retrieval of Relevant Material, Not Irrelevant Material, Enhances New Learning of a Video Lecture In-Person and Online

Zohara Assadipour [1,*], Dahwi Ahn [2] and Jason C. K. Chan [1]

[1] Department of Psychology, Iowa State University, Ames, IA 50011, USA; ckchan@iastate.edu
[2] Department of Psychology, University of Waterloo, Waterloo, ON N2L 3G1, Canada; d3ahn@uwaterloo.ca
* Correspondence: zohara@iastate.edu

Abstract

Interpolated retrieval enhances the learning of new information—a finding known as the forward testing effect. The context change account suggests that learning benefits are due to a shift in internal context, which can be triggered through the retrieval of either content-relevant or content-irrelevant information. In two experiments, we examined whether interpolated episodic, autobiographical, and semantic retrieval would enhance new learning of a video lecture, compared to interpolated review. Participants watched a STEM topic lecture divided into three ~5 min segments and completed their assigned interpolated activity after the first two segments. Across both a laboratory (Experiment 1, $N = 249$) and online setting (Experiment 2, $N = 246$), only episodic retrieval enhanced the learning of new material; autobiographical and semantic retrieval (content-irrelevant) did not improve new learning. Critically, we introduced a measure of context change to determine whether the level of engagement in these interpolated activities predicted recall. Engagement correlated with criterial test performance when controlling for effort (seriousness). Our results support a multi-factor explanation for the forward testing effect, providing evidence for both the context change and strategy change accounts, although we emphasize that support for context change should be interpreted with caution.

Keywords: learning; memory; retrieval practice; forward testing effect; context change; lecture; strategy change account; engagement; test-potentiated new learning; interpolated retrieval

1. Introduction

Students often need to learn and retain substantial amounts of information, particularly in academic settings where college lectures frequently exceed 40 min. Problematically, research has consistently shown that students' ability to learn decline over the course of extended learning sessions (Johnstone & Percival, 1976; Risko et al., 2012), and this effect is particularly pronounced in online classes (Ahn & Chan, 2022; Francis et al., 2019). To address this challenge, interpolated retrieval practice (IRP)—whereby a lesson session is broken up into smaller segments by having students answer questions about what they have learned between the segments—has been proposed as an effective intervention (Brame, 2016; Hew, 2018). Specifically, IRP enhances the retention of previously studied material (Murphy et al., 2023; Roediger & Karpicke, 2006; Rowland, 2014). But more importantly for present purposes, it also facilitates the learning of new information, a phenomenon known as the forward testing effect (FTE) or test-potentiated new learning (Chan et al., 2018b, 2020, 2025; Pastötter & Bäuml, 2014; Yang et al., 2018).

Evidence of the forward testing effect (FTE) was first reported by Tulving and Watkins (1974) using word-pairs in an A-B, A-C paired-associate learning paradigm. Specifically, participants learned two lists of pairs that shared the same cues, and the key manipulation of this study was whether or not participants completed an interpolated test for the first list (A-B) before they studied the second (A-C). In a criterial test (i.e., a final assessment of learning) that required participants to recall all targets associated with a cue (both B and C in response to A), participants recalled more of the first list (B) words than the second list (C) words if they completed an interpolated drawing task from their short-term memory. However, this effect was eliminated when participants took an interpolated test between the lists, because the interpolated test promoted the learning of items in the second list. Subsequent research has shown that engaging in interpolated retrieval practice (IRP) can insulate against the buildup of proactive interference (PI)—a negative effect of prior learning on new learning (Ahn & Chan, 2022; Pastötter et al., 2011; Szpunar et al., 2008)—and enhance list discrimination (Chan & McDermott, 2007a; Jang & Huber, 2008; Pastötter & Bäuml, 2014; Pastötter et al., 2011; Szpunar et al., 2008; Weinstein et al., 2014).

Several theoretical accounts have been proposed to explain this forward testing effect. The context change account postulates that retrieval induces an internal context change when one switches from encoding to retrieval (Davis et al., 2017; Tulving et al., 1994). For example, when participants study items in a list, those items are associated with a List-1 context. According to this account, retrieval of the List-1 items causes a change in internal context, so when participants learn new items afterwards, those items are associated with a different List-2 context. In contrast, when participants do not perform retrieval practice between the two lists, all of the list items are hypothesized to be associated with a single study-phase context. The contextual segregation between Lists 1 and 2 is believed to reduce the detrimental effects of PI and enhance the retrieval of list-specific items (Jang & Huber, 2008; Pastötter et al., 2011; Szpunar et al., 2008), thereby contributing to the occurrence of the FTE.

In addition to episodic retrieval practice, other tasks that differ substantially from episodic encoding are also believed to induce context change. These tasks include performing an N-back task (Pastötter et al., 2011), recalling details of one's childhood home, or imagining a distinctive scenario such as becoming invisible (Delaney et al., 2010; Jonker et al., 2013; Sahakyan et al., 2013; Sahakyan & Kelley, 2002). Another task that has been suggested to induce context change from episodic encoding is to perform semantic generation, in which participants produce exemplars (e.g., chair, table) based on a category cue (e.g., furniture). A commonality across the aforementioned tasks is that they involved some type of retrieval activity, even if the retrieval targets (e.g., one's childhood home or exemplars of a category) were irrelevant to the learning task. For example, the semantic generation task requires retrieval from semantic memory, the N-back task requires retrieval from working memory, and the childhood home task requires retrieval from autobiographical memory. At first glance, the imagination tasks described above do not seem to involve retrieval, but tasks that require rich imaginative constructions of scenes often engage brain regions that are critical to episodic retrieval (Schacter et al., 2017; Szpunar et al., 2007, 2009). Of particular relevance to the present study is that inserting these learning-irrelevant retrieval tasks into a multi-list learning paradigm has sometimes produced the FTE (Divis & Benjamin, 2014; Jang & Huber, 2008; Kliegl & Bäuml, 2021; Kriechbaum & Bäuml, 2024; Pastötter et al., 2008, 2011), just like a learning-relevant episodic retrieval task would. Together, these studies provided evidence for the context change account of the FTE.

However, support for the context change account is not universal. For example, Weinstein et al. (2015) did not observe an FTE using interpolated autobiographical retrieval and semantic generation, despite attempting to directly replicate Pastötter et al. (2011), but

without the EEG component. Recently, Ahn and Chan (2022) observed that interpolated episodic recall potentiated new learning to a similar degree regardless of whether participants studied inter-related (e.g., both Lists 1 and 2 contained words related to fruits and animals) or categorically distinct word lists (e.g., List 1 contained only fruit words and List 2 contained only animal words). If interpolated retrieval enhances new learning due to context change, the FTE should be much weaker or absent for categorically distinct lists. Because switching categories triggers a context change (Brown, 1958; Peterson & Peterson, 1959; Wickens et al., 1963), it should render the benefits of interpolated retrieval redundant. Therefore, the finding that the FTE was undiminished with the categorized lists relative to inter-related lists posed a problem for the context change account. Likewise, Yang et al. (2019) argued that the context change account would not predict an FTE when participants learn different types of materials across lists (e.g., List 1 contains Swahili–English pairs, and List 2 contains face–name pairs). Yet, interpolated retrieval produced similarly powerful benefits on new learning regardless of whether participants studied the same or different types of material across lists (see also Hong et al., 2019, for evidence inconsistent with the context account for the testing effect).

Some theorists have also questioned the viability of the context change account from a conceptual perspective. One concern is that the extent to which internal context change has occurred is difficult to ascertain. This issue is highlighted by prior work suggesting that interpolated activities like mental arithmetic or counting backwards by three do not induce context change (Divis & Benjamin, 2014; Kliegl & Bäuml, 2021; Pastötter et al., 2011). However, given that these interpolated activities are dissimilar to episodic encoding and are content-irrelevant to the encoded material, one might expect them to induce context change. The assumption that these tasks do not induce context change raises the following question: what makes those tasks differ from semantic generation, an N-back task, or imagining one's childhood home in terms of context change? Without an independent measure of context change, it is difficult to evaluate this claim directly. Instead, researchers have used improvement in memory performance as evidence for context change. This practice is problematic because it leads to circularity, which has been a long-standing concern for some context change explanations (Riccio et al., 1984, 1999). Context change is a *proposed explanation* for why the FTE happens, so one cannot, in turn, use the occurrence of the FTE to justify context change. In the present study, we aimed to provide a rigorous test of the context change account as an explanation of the FTE, and we introduced a way to quantify context change independently of criterial task performance.

Another prominent explanation for the FTE is the strategy change account, according to which, engaging in retrieval practice promotes future learning because it encourages individuals to apply more effective encoding and/or retrieval strategies later (Chan et al., 2018a, 2018b; Davis & Chan, 2015, 2023). Specifically, taking a test gives the learner metacognitive insights by revealing the test structure, retrieval cue availability, gaps in knowledge, etc., which can inform future learning (Bjork & Storm, 2011; deWinstanley & Bjork, 2004; Hays et al., 2013). For example, after taking a test, participants might be underwhelmed by their performance and attempt to exert a greater effort into learning subsequent material, as demonstrated by increased study time (Davis & Chan, 2023; Soderstrom & Bjork, 2014; Yang et al., 2017). Alternatively, during an interpolated test, participants might learn that the relational properties of studied items can serve as powerful retrieval cues. This realization might cause participants to focus on these properties during subsequent encoding and retrieval opportunities. This type of strategy change can manifest via greater organization in participants' recall (Chan et al., 2020, 2018a; Jing et al., 2016; Yang et al., 2022). However, the strategy change perspective cannot readily explain why content-irrelevant retrieval tasks, like semantic generation, would promote new learning. Bäuml and colleagues (Kliegl

& Bäuml, 2021, 2023) thus proposed that combining strategy change and context change into a two-factor account can address each of the individual account's shortcomings. Indeed, as other theorists have also noted (Chan et al., 2018b; Yang et al., 2018), the strategy change and context change accounts are complementary rather than mutually exclusive, so combining them can enhance explanatory power without posing conceptual contradictions.

2. The Current Study

There is now a considerable body of literature demonstrating the robust nature of the FTE; however, the vast majority of these studies have employed simplistic materials (e.g., word lists, paired associates) that students do not typically encounter in the physical or virtual classroom (for a review, see Chan et al., 2018b). In the current study, we sought to examine whether content-irrelevant IRP can enhance the learning of lecture materials, which remains the most common form of knowledge delivery in both the classroom and online (Hansch et al., 2015; Seaton et al., 2014). To that end, we examined the impact of different types of IRP on the new learning of a video lecture under both in-person (Experiment 1) and online settings (Experiment 2). The strategy change and context change accounts make distinct predictions regarding the effects of different types of IRP on new learning.

In the present study, participants watched a video lecture split into three segments. Immediately following Segments 1 and 2, participants either studied several *review* slides, performed *episodic retrieval* by answering six quiz questions about the immediately prior segment, performed *autobiographical retrieval* by thinking about their childhood home or a vacation, or performed *semantic retrieval* by generating exemplars from category names.

Both the strategy change and context change accounts would lead one to predict that interpolated episodic retrieval would promote new learning—as shown by an increase in the criterial Segment 3 test performance—relative to an interpolated review. We were less certain about whether interpolated autobiographical retrieval and semantic retrieval would promote new learning. On the one hand, the strategy change account predicts that only episodic retrieval (but not semantic or autobiographical retrieval) should potentiate new learning; on the other hand, the context change account might lead one to predict that all interpolated retrieval tasks would promote new learning.

Prior research has employed autobiographical retrieval and semantic generation as methods to induce context change. However, heretofore, no studies have attempted to measure the extent of context change that participants experienced. Indeed, none have reported participants' performance during these interpolated tasks (Divis & Benjamin, 2014; Kliegl & Bäuml, 2021, 2023). In the present paper, we measured participants' engagement and success with the interpolated activities. Assuming that mental context change is a continuous rather than a binary phenomenon, participants who are more engaged with the interpolated tasks should undergo greater context change. To this end, we introduce a measure for both autobiographical retrieval and semantic generation. Engagement with the autobiographical retrieval task was measured via a word count after verification that participants were properly following instructions (e.g., they were recalling their childhood home or a vacation rather than something unrelated), and semantic retrieval engagement was measured by the number of correct exemplars recalled.

If mental context change is a driver of the FTE, then one should expect a positive association between engagement with these tasks and criterial test performance. However, before accepting this hypothesis at face value, it must be weighed against an alternative explanation, namely, that participants who engage more with autobiographical or semantic retrieval may exhibit better memory performance not because of context change, but because they were more engaged with the experiment as a whole. In other words, these

participants were simply taking the experiment more seriously. To address this alternative hypothesis, we also asked participants to indicate how seriously they took the experiment using a Likert scale question. We report a hierarchical regression analysis to examine the extent to which interpolated task engagement predicts criterial test performance beyond seriousness. Moreover, we detail limitations regarding the interpretation of these data as a proxy for context change in Section 6.

3. Experiment 1: Laboratory Study

3.1. Methods

3.1.1. Design and Participants

This study adopted a between-subjects design wherein participants were randomly assigned to one of four interpolated activity conditions: (content-relevant) episodic retrieval, (content-irrelevant) autobiographical retrieval, (content-irrelevant) semantic retrieval, or (content-relevant) review. Target sample size was determined based on Pastötter et al. (2011), who reported a minimal effect size of $d = 0.70$ across all interpolated retrieval tasks (content-relevant episodic retrieval, semantic generation, working memory) relative to the control condition of interpolated review or interpolated distractor. To be conservative, we conducted our power analysis using $d = 0.49$, which is 70% of Pastötter et al.'s effect size. When there are more than two levels in the independent variable, the analysis of variance (ANOVA) approach is commonly used for the power analysis. However, because we were examining against the hypothesis that performance on the Segment 3 criterial test would be equivalent in the three retrieval conditions, we used an independent samples *t*-test approach with an unequal allocation ratio (i.e., retrieval/review = 3:1). This analysis indicated that we would need 178 participants (133 in the retrieval conditions and 45 in the review condition) to achieve 80% power at a 0.05 alpha level.

Participants consisted of 249 students (M_{age} = 19.06, $N_{reported}$ = 226) from a large midwestern university and were randomly assigned to the four conditions. There were 61 participants in autobiographical retrieval, 62 in semantic retrieval, 63 in episodic retrieval, and 63 in review. The experiment was conducted in the laboratory under the supervision of research assistants. Participants received course credit as compensation. See Table 1 for participant demographics.

Table 1. Experiment 1 participant demographic information.

Race	*N*
Asian/Asian American	20
Black/African American	7
Hispanic/Latinx	17
Multi-racial	2
Other	2
White	207
Decline to disclose	3
Gender	***N***
Man	80
Non-binary	2
Woman	166
Decline to disclose	1

Note. Participants were able to report multiple races.

3.1.2. Materials and Procedure

Participants watched a ~15 min lecture video, which was divided into three ~5 min segments. The lecture covered topics in either statistics or physics. For stimulus sampling

purposes, two versions of each lecture were created featuring either a man or woman instructor. Each lecture displayed the instructor in a head-and-shoulder view at the bottom left of the lecture slides. Details about the construction of the lecture videos can be found in Chan et al. (2025).

Participants were instructed to pay attention to the lecture like they would for an actual class, but they were prohibited from taking notes. Participants were told that a randomization algorithm determined whether they would receive quiz questions or review slides following each lecture segment. However, in actuality, participants engaged in their assigned interpolated activity for three minutes following each of the first two segments. All participants took a criterial test of Segment 3 immediately following its presentation.

In the episodic retrieval condition, participants answered six short-answer questions over the just-presented lecture segment (for Segments 1 and 2). For example, a question from the physics lecture asked, "What type of particles do masers emit?" Questions were presented individually for 30 s without feedback, so the episodic retrieval task lasted a total of three minutes.

In the review condition, participants were shown six review slides over content presented in the previous lecture segment for Segments 1 and 2, with each shown for 30 s. All review slides had equivalent information to the quiz questions in the episodic retrieval condition, such as "Masers emit microwaves and radio waves."

Participants in the autobiographical retrieval condition were prompted to recall a specific memory as extensively as possible for three minutes. Specifically, they were asked to recall their childhood home following one segment and their most memorable vacation following another segment, with the order of prompts counterbalanced across participants. Participants were warned, in red and underlined letters, not to include any personally identifiable information during their recall. No participants included identifying information.

In the semantic retrieval condition, participants were instructed to recall as many exemplars as possible for a given category. The categories included four-footed animals, furniture, non-alcoholic beverages, professions, sports and vegetables. Participants were given 60 s per category and were shown a randomly chosen, non-repeated three categories following each of the first two lecture segments.

Following Segment 3, all participants took a criterial test on this segment. This test consisted of six short-answer questions and was administered in the same manner as the interpolated episodic retrieval task. Before dismissal, all participants answered demographic and data quality control questions.

3.2. Results

All analyses were conducted using two-tailed tests with an alpha level of 0.05. When reporting effect sizes, we use Cohen's *d* for pairwise comparisons and partial eta square η_p^2 for other comparisons. All analyses were conducted with JASP Team (2024). We report the results regarding context change based on interpolated task engagement following presentation of the results from both experiments. These analyses were based on the combined data from Experiments 1 and 2 because our individual experiments were not powered to detect the association between interpolated task engagement and test performance. All inferential statistics are reported to three decimal places, and descriptive statistics are reported to two decimal places.

Do All Interpolated Retrieval Tasks Boost New Learning?

The primary question of the present study is to compare the effectiveness of different types of retrieval tasks on new learning. Figure 1 shows the criterial test performance across

the four interpolated activities. The context change account posits that retrieval would induce a change in context and reduce proactive interference, so all retrieval conditions should outperform review. An ANOVA was conducted to examine this hypothesis with the independent variable consisting of each interpolated activity condition and the dependent variable as the proportion of correct recall on the criterial test. Results showed a significant main effect of interpolated activity, $F(3, 245) = 4.96$, $p = 0.002$, $\eta_p^2 = 0.057$. Perhaps most importantly, planned comparisons showed that only interpolated episodic retrieval ($M = 0.60$, $SD = 0.27$) significantly enhanced new learning relative to interpolated review ($M = 0.44$, $SD = 0.22$), $t(124) = 3.61$, $p < 0.001$, $d = 0.642$. Neither autobiographical retrieval ($M = 0.52$, $SD = 0.25$), $t(122) = 1.78$, $p = 0.078$, $d = 0.319$, nor semantic retrieval ($M = 0.47$, $SD = 0.26$), $t(123) = -0.53$, $p = 0.594$, $d = -0.096$, significantly promoted new learning relative to the review. Additionally, lecture topics did not interact with interpolated activity, $F(3, 241) = 0.97$, $p = 0.410$, $\eta_p^2 = 0.012$. Our results are thus consistent with the strategy change account but not the context change account. Descriptive statistics of each condition are presented in Table 1. One might notice, however, that the effect for autobiographical retrieval was somewhat inconclusive (i.e., its p-value was between 0.05 and 0.10). We will address this finding in more detail following presentation of the data in Experiment 2, which will serve as a replication in an online setting.

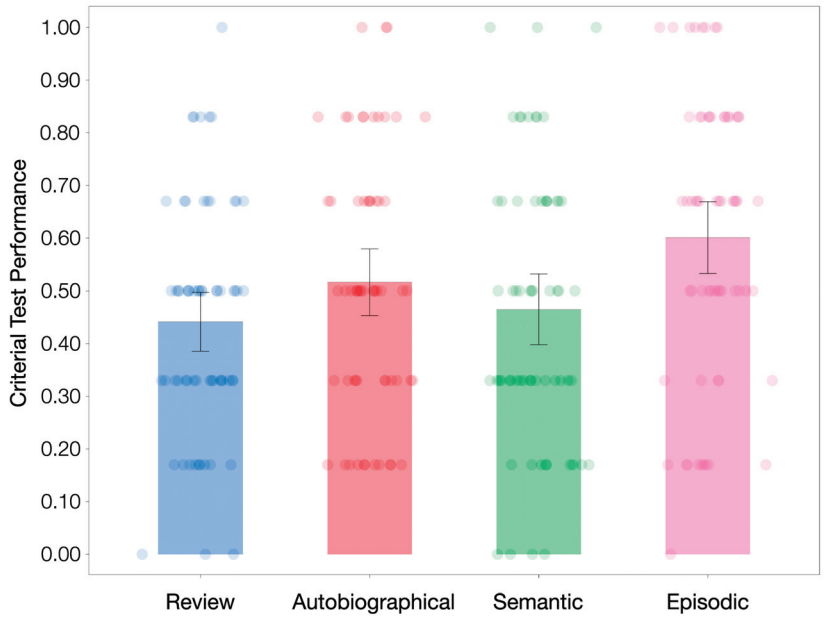


Figure 1. Criterial test performance across conditions of Experiment 1. Dots denote individual participant performance. Error bars denote 95% descriptive confidence interval. Direct comparisons reveal that episodic retrieval, but not semantic retrieval, significantly enhances new learning relative to review. The difference between autobiographical retrieval and review is inconclusive ($p = 0.078$).

4. Experiment 2: Online Study

4.1. Methods

4.1.1. Design and Participants

Experiment 2 was a direct replication of Experiment 1, except that we wanted to generalize the data to online learning. Consequently, all participants completed the experiment fully online. Participants consisted of both students from the same midwestern university as in Experiment 1 ($N = 49$) and those from Prolific ($N = 200$). We began data collection near the end of spring semester, during which university participants were available, but we finished data collection with participants from Prolific after the end of the spring semester. The university students were compensated with course credit, whereas Prolific participants

were compensated with USD 4. We excluded data from two university students who did not take the experiment seriously at all (i.e., reporting a 1 on a 6-point scale). One additional participant from Prolific was excluded for not following instructions during the autobiographical retrieval task (the person attempted to recall details of the lecture instead). This resulted in a total of 246 participants (M_{age} = 35.80, $N_{reported}$ = 245) in all analyses. Participants were again randomly assigned to the four interpolated activities: 61 completed autobiographical retrieval, 68 semantic retrieval, 56 episodic retrieval, and 61 review. The sample size was slightly uneven across the conditions because of random assignment. See Table 2 for participant demographics.

Table 2. Experiment 2 participant demographic information.

Race	***N***
Asian/Asian American	21
Black/African American	50
Hispanic/Latinx	22
Multi-racial	7
Other	4
White	148
Decline to disclose	4
Gender	***N***
Agender	1
Man	121
Non-binary	4
Woman	115
Transgender	2
Decline to disclose	3

Note. Participants were able to report multiple races.

4.1.2. Materials and Procedure

The same materials and procedure were adopted from Experiment 1, except that all participants completed their experiment online.

4.2. Results

Only Content-Relevant Episodic Retrieval Enhanced New Learning

Figure 2 shows the criterial test performance per condition. Similar to Experiment 1, an ANOVA showed a main effect of interpolated activity on criterial test performance, $F(3, 242) = 3.24$, $p = 0.023$, $\eta_p^2 = 0.039$. More importantly, only episodic retrieval ($M = 0.55$, $SD = 0.26$) enhanced new learning relative to review ($M = 0.46$, $SD = 0.25$), $t(115) = 2.04$, $p = 0.044$, $d = 0.377$. Once again, neither semantic retrieval ($M = 0.41$, $SD = 0.28$), $t(127) = -1.00$, $p = 0.317$, $d = -0.177$, nor autobiographical retrieval ($M = 0.45$, $SD = 0.24$), $t(120) = -0.19$, $p = 0.854$, $d = -0.033$, resulted in significantly better criterial test performance than review. If anything, both effects were in the negative direction. The data in the autobiographical retrieval condition are particularly noteworthy, given that participants in this condition showed marginally better criterial test performance than those in the review condition in Experiment 1. The complete absence of the effect here suggests that the marginal effect in Experiment 1 was perhaps unreliable. Similar to Experiment 1, lecture topics did not interact with interpolated activity, $F(3, 238) = 0.19$, $p = 0.906$, $\eta_p^2 = 0.002$.

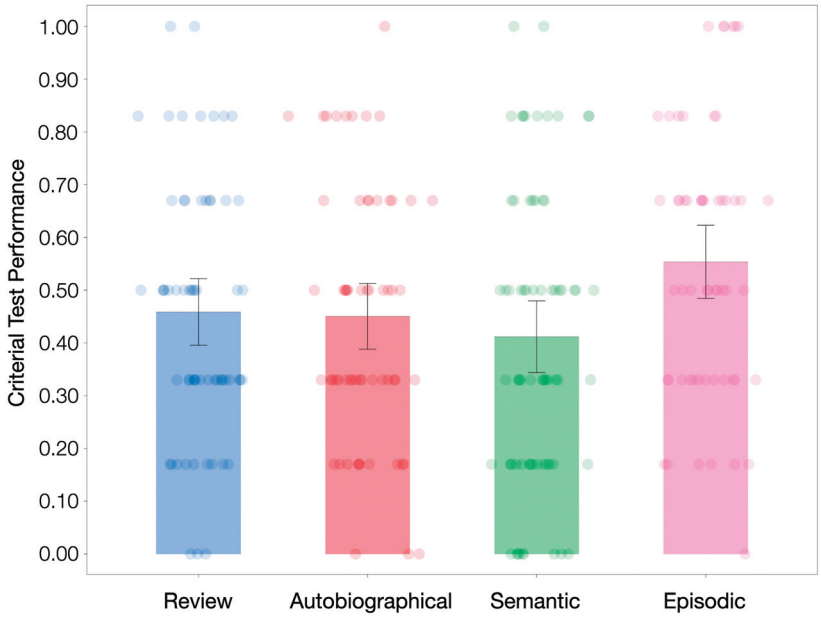


Figure 2. Criterial test performance across conditions of Experiment 2. Error bars denote 95% descriptive confidence interval. Direct comparisons reveal that episodic retrieval, but not semantic or autobiographical retrieval, significantly enhances new learning when compared to review.

5. Exploratory Analyses

In the following, we report several exploratory analyses that used the combined data from Experiments 1 and 2 to increase statistical power. First, one might wonder if the inconclusive result of autobiographical retrieval on new learning in Experiment 1 was due to insufficient power. To address this concern, we report an exploratory analysis by combining the data from the two experiments. Next, we report a series of analyses aimed at quantifying context change and its contribution to criterial recall performance. We did not conduct these analyses using data from the individual experiments because they were not powered for these exploratory analyses.

5.1. Lecture Relevance as a Determinant of Test-Potentiated New Learning

An ANOVA with the interpolated activity and experiment (1 vs. 2) as independent variables and criterial test performance as the dependent variable showed a main effect for interpolated activity, $F(3, 487) = 7.41$, $p < 0.001$, $\eta_p^2 = 0.044$. Neither the main effect of the experiment, $F(1, 487) = 2.64$, $p = 0.105$, $\eta_p^2 = 0.005$, nor the interaction, $F(3, 487) = 0.66$, $p = 0.580$, $\eta_p^2 = 0.004$, was significant. Most importantly, similar to the conclusions from the individual experiments, interpolated episodic recall of content-relevant content promoted new learning, $t(241) = 4.02$, $p < 0.001$, $d = 0.516$, whereas autobiographical and semantic recall did not, $t_{autobiographical}(244) = 1.09$, $p = 0.275$, $d = 0.139$; $t_{semantic}(252) = 0.41$, $p = 0.686$, $d = 0.051$, despite the increase in statistical power of the combined data set. The null effect for interpolated autobiographical retrieval is particularly important, as the result in Experiment 1 was somewhat inconclusive. We now report data that bear relevance to the context change perspective.

5.2. Measuring Context Change and Its Impact on New Learning

As mentioned in the Introduction, we aimed to use interpolated task engagement as a measure of context change. If context change enhances new learning, one might expect a positive correlation between interpolated task engagement and criterial test performance. To ensure that any positive association between these measures was not simply driven by a general increase in effort, we factored out participants' reports of seriousness in a

hierarchical regression analysis. Specifically, the first model regressed seriousness on criterial test performance, and the second model regressed both seriousness and interpolated task engagement on criterial test performance. To quantify interpolated task engagement for autobiographical retrieval, we measured the total number of words that participants recalled during the two autobiographical retrieval trials. Two research assistants screened the autobiographical recall records and found that all participants were on-task (i.e., recalling what they were asked). To measure interpolated task engagement for semantic retrieval, we counted the number of correct exemplars recalled per participant based on Van Overschelde et al.'s (2004) category norms. To avoid double counting exemplars, we only included unique compound terms (e.g., Coke and Coca-Cola) but not repeated compound terms (e.g., beans, but not green beans and string beans).

We first report data for participants in the *episodic retrieval* condition. Here, we expect that interpolated recall would be positively associated with both criterial recall and seriousness ($M = 4.99$, $SD = 1.00$). Indeed, the regression model for seriousness was significant, $F(1, 117) = 9.55$, $p = 0.002$, $r = 0.28$. The second model, which included seriousness as well as recall probability of Segment 1 ($M = 0.63$, $SD = 0.27$) and Segment 2 ($M = 0.63$, $SD = 0.30$), was also significant, $F(3, 115) = 11.22$, $p < 0.001$, $r = 0.48$. Critically, the addition of Segments 1 and 2 recall probability significantly improve the model, $r^2_{\text{change}} = 0.15$, $F_{\text{change}} = 11.22$, $p < 0.001$. Consequently, interpolated recall success predicted criterial recall performance beyond the contribution of overall task effort as measured by seriousness.

We now report data for participants in the other conditions. We first examine the data for participants in the *review* condition. Unlike participants in the episodic retrieval condition, seriousness ($M = 4.98$, $SD = 1.01$) was *not* significantly associated with criterial recall performance, $F(1, 122) = 1.74$, $p = 0.189$, $r = 0.12$. Because participants in the review condition did not complete any interpolated tasks that produced measurable data, we did not complete a hierarchical regression for them.

For participants in the *autobiographical retrieval* condition, the first model showed that seriousness ($M = 5.00$, $SD = 1.02$) was weakly and positively associated with criterial test performance, $F(1, 119) = 4.11$, $p = 0.045$, $r = 0.18$. The second model, with the total word count of autobiographical retrieval as an additional predictor ($M = 214.99$ words, $SD = 80.27$ words), was also significant, $F(2, 118) = 4.98$, $p = 0.008$, $r = 0.28$. Most importantly, autobiographical recall significantly increased the predictive power of the regression model, $r^2_{\text{change}} = 0.04$, $F_{\text{change}} = 5.69$, $p = 0.019$. Specifically, participants who were more engaged with the autobiographical recall task (as shown through a greater word count) showed better criterial test performance, and this association accounted for unique variance beyond overall task engagement as measured by seriousness. See Panel (a) of Figure 3 for a scatterplot.

Participants in the *semantic retrieval* condition exhibited a similar pattern to those in the autobiographical retrieval condition. Specifically, the first model revealed a marginal positive association between seriousness ($M = 4.95$, $SD = 1.01$) and criterial recall performance, $F(1, 128) = 2.83$, $p = 0.095$, $r = 0.15$. The second model, with the addition of the total number of correct exemplars generated across six categories ($M = 51.92$, $SD = 13.07$), was strongly predictive of criterial recall performance, $F(2, 127) = 11.88$, $p < 0.001$, $r = 0.40$. See Panel (b) of Figure 3 for a visual representation. Indeed, semantic retrieval success contributed substantially to the model, $r^2_{\text{change}} = 0.14$, $F_{\text{change}} = 20.49$, $p < 0.001$. Together, the results from the semantic and autobiographical retrieval conditions show that, despite not conferring an overall benefit to new learning, participants who were more engaged with the interpolated tasks performed better on the criterial test. This outcome is consistent with the context change account.

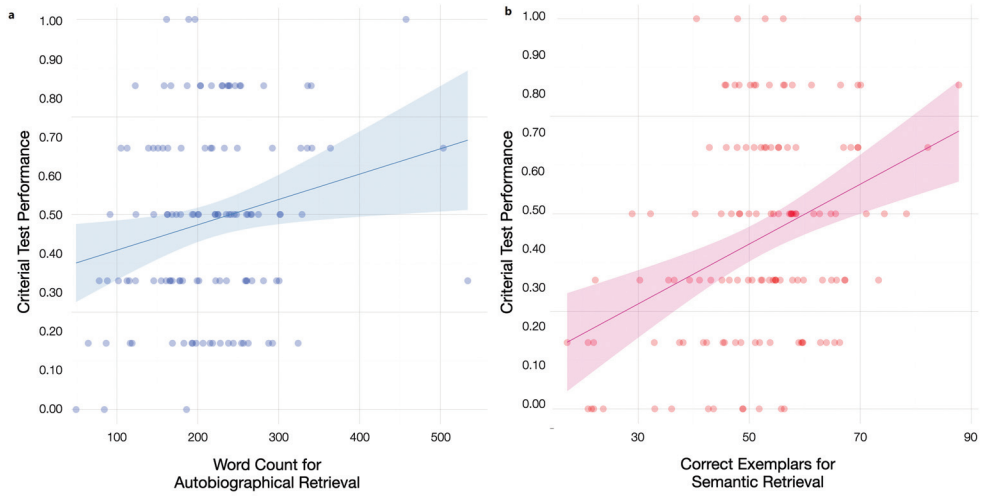


Figure 3. Relationship between engagement in content-irrelevant retrieval conditions and criterial test performance. Horizontal jitter is introduced to improve data visibility. We scale the amount of jitter to the content-irrelevant retrieval variable. The scatterplot in Panel (**a**) shows autobiographical retrieval engagement and criterial test performance. Engagement was measured by word count (horizontal jitter ±5). The scatterplot in Panel (**b**) shows semantic retrieval engagement and criterial test performance. Engagement is measured by the number of correct exemplars generated (horizontal jitter ±1).

6. Discussion

In two experiments, we consistently found that the interpolated episodic retrieval of lecture content promoted new learning of the lecture relative to the review. However, content-irrelevant interpolated retrieval, such as recalling a family vacation and childhood home or generating exemplars from category names, failed to promote new learning. This pair of experiments was the first to show that answering interpolated quiz questions can enhance learning of a video lecture in both laboratory and fully online settings. We now discuss their applied and theoretical implications.

6.1. Interpolated Retrieval Online and In-Lab

Although considerable research has demonstrated the benefits of interpolated quizzing on new learning (Chan et al., 2018a; Pastötter & Bäuml, 2014; Yang et al., 2018), most of this work employed experimental material that is unlikely to be encountered by students in the classroom. The contrived materials, such as word lists and paired associates, are advantages for theoretical investigations because they allow researchers to examine the data at a fine-grained level. However, they also limit the data's generality. A few studies have demonstrated the benefits of interpolated retrieval on new learning with realistic lecture content, but, until very recently (Chan et al., 2025), all of these have tested participants in the laboratory, despite using the term "online lecture" to describe their material (e.g., Conrad & Newman, 2021; Pan et al., 2020; Szpunar et al., 2013). Here, Experiment 1's participants were tested in the laboratory, and Experiment 2's participants were tested online with the same materials and procedure, and we found that (i) answering interpolated quiz questions promoted new learning both in the laboratory (Experiment 1) and online (Experiment 2), and (ii) the benefits of interpolated retrieval appear smaller in Experiment 2 (d = 0.38) than in Experiment 1 (d = 0.64). Note that, when we compared the effect sizes in an ANOVA, the interaction between interpolated testing and the learning environment across the experiments was not significant, F = 1.00, p = 0.319. Nonetheless, our study was also not powered to detect this interaction.

Notably, the moderate interpolated retrieval benefit in our Experiment 2 mirrored that reported by Chan et al. (2025, d = 0.37, N = 235), who tested all of their participants online. Chan et al. (2025) suggested that the modest effect size in their experiment might be the result of online participants being distracted when attempting to learn the lecture material. In particular, they argued that typical in-lab studies present participants with a pristine learning environment, where they are not allowed to use mobile phones, browse the internet, talk to others, or engage with other distractions. The prohibition of these activities fosters a nearly ideal learning environment that diverges considerably from how students typically learn online (Adhani & Remijn, 2023; Aivaz & Teodorescu, 2022; Hollister et al., 2022). Consequently, the small FTE might reflect the level of benefits that one can expect from interpolated testing in distraction-filled online learning scenarios. Although this proposal is reasonable, Chan et al. (2025) did not test their lecture materials in the laboratory, so the small effect size could be a consequence of the materials rather than online learning being more distracting per se.

In the present study, the same lecture material produced non-significantly different effect sizes based on testing environments, so the lecture material is unlikely to be the culprit for the small (and nearly identical) effects exhibited in our Experiment 2 and in Chan et al. (2025). Rather, we tentatively conclude that online learning environments, which are often more distracting than in-lab studies, might diminish the benefits of interpolated testing. We emphasize that this hypothesis is tentative because it involves a cross-experimental comparison and different participant populations. Specifically, participants in Experiment 1 were university students, and participants in Experiment 2 featured both university students and Prolific participants. Even if we had enrolled only university students in Experiment 2, the participant population would still be confounded across experiments because those who self-select into in-person and online studies might have different characteristics. To fully remove this potential subject-selection confound, one must conduct a study in which participants are recruited for either an in-person or online study and are then randomly assigned to one of the environments after they have arrived at the laboratory. Specifically, participants cannot know where (in-lab or online) they would complete the experiment when they sign up or show up for the study. We are not aware of such a study at the current time, so a definitive answer to the question about online vs. in-person learning awaits further research. Moreover, we consider that Experiment 2 closely resembles the real-life learning contexts for online and asynchronous courses. At the very least, the present study showed that answering interpolated quiz questions, but not interpolated irrelevant retrieval activities, can promote the new learning of a video lecture both in-person and online.

6.2. Interpolated Retrieval, Strategy Change, and Context Change

In two experiments, we found that interpolated episodic retrieval enhanced subsequent learning, whereas interpolated semantic and autobiographical retrieval did not. The latter results have important theoretical and applied implications. Although researchers have sometimes reported that semantic recall can facilitate new learning (Divis & Benjamin, 2014; Pastötter et al., 2011), the effect was not always found (Weinstein et al., 2015). Recent studies have suggested that semantic generation potentiates new learning (Kliegl & Bäuml, 2021, 2023; Kriechbaum & Bäuml, 2024) only when the pre- and post-interpolated retrieval materials are unrelated to each other (e.g., when participants studied lists of noncategorized words), but not when they are related to each other (e.g., when participants studied inter-related word lists). Bäuml and colleagues explained the dissociation as supporting the context change account—specifically, performing semantic retrieval between the episodic encoding of word lists causes a mental context change that facilitates list discrimination.

For example, after studying List 1, which is associated with a List 1 context, semantic generation alters the mental context from encoding to retrieval. When participants then study List 2, these words are encoded under a new, List 2 context. Later, when participants attempt to recall the words from only List 2, they can use the isolated List 2 context as a retrieval cue, thereby constraining retrieval candidates (Shimizu & Jacoby, 2005). This context change account assumes that, when participants do not perform retrieval between the noncategorized lists—such as by practicing mental arithmetic or restudy—the context between Lists 1 and 2 does not change. Therefore, when participants attempt to retrieve List 2 items later, a single study context is associated with both List 1 and 2 words, making it more difficult to achieve accurate recall and minimize intrusions (for evidence that reduced interference might not contribute to the forward testing effect, see Ahn & Chan, 2022; Boustani et al., 2023).

Why then would semantic generation fail to promote new learning for related items? According to Bäuml and colleagues, when the post-retrieval List 2 contains items from categories in a pre-retrieval List 1, the List 2 items remind participants of List 1 and reinstate the pre-retrieval context (Kliegl & Bäuml, 2021, 2023; Kriechbaum & Bäuml, 2024). Consequently, even if interpolated semantic generation changes participants' mental context, the items in List 2 bring participants back to their List 1 context, thus making the retrieval of List 2 items as difficult as when participants did not perform an interpolated task that changes context (e.g., mental arithmetic, restudy). Critically, Bäuml and colleagues' two-factor account attributes the benefits of episodic retrieval for the new learning of categorized items not to context change, but to strategy optimization. That is, when participants perform interpolated episodic retrieval, not only does it induce a mental context change, but the task also provides participants with valuable information about their own learning, and participants can use this metacognitive knowledge to optimize their subsequent encoding and retrieval (Ahn & Chan, 2024; Chan et al., 2020, 2018a). For example, participants might realize that they are underperforming their expectations and therefore exert greater time or effort into subsequent learning (Davis & Chan, 2023; Soderstrom & Bjork, 2014; Yang et al., 2017). Moreover, prior retrieval practice can promote memory organization during subsequent recall (Chan et al., 2020; Zaromb & Roediger, 2010).

In the present study, we found that neither semantic generation nor autobiographical recall promoted new learning of the lecture. This outcome is consistent with both the strategy-change account alone and the two-factor account. From the strategy-change perspective, content-irrelevant retrieval tasks should not help participants optimize their subsequent learning. From the two-factor account's perspective, because the pre- and post-retrieval segments covered the same topic, the later segments might have reinstated the context of the prior segments and negated any context change benefits.

A key differentiator between these two accounts, however, is that we provided the first demonstration that greater context change, as quantified by interpolated task engagement, was associated with better criterial test performance, even when the overall task (i.e., autobiographical and semantic retrieval) did not produce a forward testing effect. This association is difficult to reconcile if one subscribes to the view that strategy change is the only contributor to the FTE, but it is consistent with the two-factor account.

Although the present results support the two-factor account, we continue to urge researchers to use caution when employing mental context change—which is unobservable—as an explanatory mechanism. We believe that the account of context change must satisfy two criteria for it to be truly useful: First, context change needs to be quantifiable independently of the criterial measure. Second, researchers should delineate why some tasks induce context change and others do not.

Regarding the first criterion of quantifying context change, one should not use better memory performance—which one is attempting to explain via context change—as a proxy for context change having occurred, because doing so creates a circular logic. The present study partly addressed this concern by providing a measure of context change during which participants engaged with the interpolated activity. But we emphasize that interpolated task engagement is only one possible way to assess context change, and our effort represents an early, preliminary attempt to measure it. In fact, even with these measurements, one cannot ascertain that a mental context change has occurred. Instead, other factors such as verbal fluency (which can affect both semantic generation and recall, Chan & McDermott, 2007b; Glisky et al., 1995) might underlie our measure of interpolated task engagement. Moreover, although Bauml and colleagues' version of the context change account can explain why content-irrelevant retrieval tasks can promote new learning for unrelated materials but not related materials, this account requires one to accept *two* assumptions: first, that context change has occurred during the interpolated activity, and second, that related materials reinstate the pre-retrieval context. In this study, we sought to quantify the first assumption. Attempting to provide independent verification of the second assumption might prove even more difficult.

Concerning the second criterion, researchers should specify why some tasks (e.g., episodic and semantic retrieval) induce context change and others do not (e.g., mental arithmetic, drawing copy, restudy). Alternatively, researchers should provide careful conceptual analyses on the influence of a task on performance, without resorting to assumptions about whether a task causes context change. Some researchers have suggested that retrieval tasks (semantic, autobiographical, working memory, episodic) can induce context change relative to encoding, whereas tasks that do not involve retrieval (e.g., math) would not (Kliegl & Bäuml, 2021; Pastötter et al., 2011). We find this argument problematic because context change is meant to explain why retrieval potentiates new learning, so equating context change with retrieval again risks circularity. Further, this assumption fails to answer the question of why retrieval induces context change, whereas other tasks like math or drawing do not, even though these latter tasks seem just as different from episodic encoding. Indeed, in Tulving and Watkins' (1974) interpolated drawing task, participants were shown a picture for 15 s and were then tasked to reproduce the picture by drawing it from short-term memory (similar to N-back). So it is unclear why this task would not enhance new learning from a context change perspective.

Bäuml and colleagues' explanation for why semantic retrieval does not promote new learning for categorized material, however, might provide a promising way forward (Kliegl & Bäuml, 2021, 2023; Kriechbaum & Bäuml, 2024). Specifically, rather than focusing on whether or not a task would cause a context change, one might instead consider the contribution of multiple components of a task together. Applying this type of task analysis to the finding that interpolated mental arithmetic fails to promote new learning, the important question is perhaps not whether or why math does not cause a context change. Indeed, we believe that mental arithmetic would induce a context change from episodic encoding. Rather, it is possible that mental arithmetic is a difficult task for most participants, and they either experience minor mental fatigue from the task or develop disinterest in the task, which carries over to subsequent learning. If this were the case, one might expect that, unlike autobiographical recall (a self-relevant task) or semantic generation (an interesting and challenging task), greater task engagement for mental arithmetic would translate to poorer, instead of better, criterial test performance. To adequately test this idea, however, researchers need to control for participants' math ability by giving those who are better at math more difficult questions. There are two advantages to this task analysis approach: First, researchers no longer need to explain why some tasks might

induce a context change whereas other tasks might not. Second, one can produce testable hypotheses about the relationship between interpolated task engagement and new learning on a task-by-task basis.

Lastly, we would be remiss not to mention that not all empirical evidence has been consistent with the two-factor strategy and context account. For example, according to Bäuml and colleagues, interpolated non-episodic retrieval tasks should promote new learning when the pre- and post-retrieval items are not related to each other (Kliegl & Bäuml, 2021, 2023; Kriechbaum & Bäuml, 2024). However, Weinstein et al. (2015) showed that interpolated autobiographical and semantic retrieval failed to promote new learning, whereas episodic retrieval did, regardless of whether participants studied related or unrelated words across lists. Further research is needed to clarify support for the two-factor account.

6.3. Limitations and Constraints on Generality

In the present study, we defined content relevance as the extent to which the to-be-recalled content was related to the lecture material. Consequently, the episodic retrieval condition, in which participants were asked specific short-answer questions about the lecture, was considered content-relevant. In contrast, the autobiographical and semantic retrieval conditions were not content-relevant. However, we have not included an episodic retrieval task that asked participants short-answer questions unrelated to the just-lectured content. For example, imagine a student is attending a lecture on microbiology but has an exam on cognitive psychology during the next class period. The student might attempt to recall specific facts about cognitive psychology at various points during the microbiology lecture. Might such content-irrelevant retrieval enhance new learning of the microbiology lecture? Recent research suggests that the answer might be yes. For example, Yang et al. (2019) showed that the forward testing effect is transferable across different types of learning material. In one experiment, participants studied three successive sets of individual written statements about famous artists and then learned a set of painting-artist pairs. Following each of the three sets of written statements, participants either restudied the statements or retrieved them via fill-in-the-blank questions. Afterwards, participants learned a different type of material, which was a set of painting-artist pairs, and took a criterial test on it. A forward testing effect was found in this criterial test (see also Lee & Ahn, 2018). Therefore, recalling previously studied material can potentiate new learning of an unrelated set of new material.

When applying this logic to the example scenario above, it is reasonable to predict that recalling facts about cognitive psychology might facilitate the learning of microbiology. However, a critical difference between the Yang et al. (2019) finding and the content-irrelevant episodic retrieval example above is the following: In Yang et al., participants performed interpolated retrieval on the just-studied material. In contrast, in the content-irrelevant episodic retrieval example, the hypothetical student performs interpolated retrieval on other, distally studied material. However, because the forward testing effect transfers across domains and across a lag (Chan et al., 2018a; Kliegl & Bäuml, 2021), we tentatively predict that interpolated testing would enhance new learning even when the retrieved topic was not just-studied.

We also acknowledge the limitations in the generality of our findings, as we cannot extend them to longer lectures (e.g., 1–2 h) or semester-long learning. Future research is needed to test these possibilities. Further, should testing of a distal topic enhance new learning, it is important to determine the influence of topic proximity (in terms of the lag between initial learning and interpolated testing) on the magnitude of the forward testing benefit. Lastly, our data were collected with university students and participants from the Prolific online participant pool, and, although we included two lecture topics (statistics

and physics), these materials do not resemble other types of materials that college students experience (e.g., group work, math, and art). Therefore, we urge caution when attempting to generalize these findings beyond the confines of our study.

7. Conclusions

The present experiments demonstrate that engaging in content-relevant episodic retrieval can reliably enhance new learning in both laboratory and online settings. Crucially, no benefits to new learning emerged from content-irrelevant autobiographical and semantic retrieval tasks among our stimuli and samples. These findings further show the robustness of the interpolated retrieval as an educational intervention, but they also highlight that retrieval alone is not sufficient to enhance new learning for some situations. Rather, it should be content-relevant. This outcome contrasts with previous findings that employed content-irrelevant retrieval among simplistic materials (Divis & Benjamin, 2014; Kliegl & Bäuml, 2021; Pastötter et al., 2011). Educators and students are encouraged to incorporate content-relevant episodic retrieval to promote learning. Educators may easily implement interpolated quizzes both during in-person lectures and online through short quizzes or clicker questions, requiring each student to engage in content-relevant episodic retrieval.

Our findings align with the idea that both strategy and context change contribute to the forward testing effect (Chan et al., 2018a; Kliegl & Bäuml, 2021; Yang et al., 2018). Engaging in episodic retrieval induces a context change and promotes strategy optimization, whereas autobiographical and semantic retrieval do not provide these same benefits. Consistent with Kliegl and Bäuml's (2021) context change explanation, the reinstatement of the initial lecture context likely negates any context change benefits that content-irrelevant retrieval may offer. Additionally, the content-irrelevant retrieval tasks do not offer strategic benefits (e.g., potential gaps in knowledge, retrieval cues, and structure of criterion test), where episodic retrieval would. Thus, the two-factor account predicts that autobiographical and semantic retrieval would not produce the forward testing effect in instances where the learned materials are semantically related, as was the case in the presented study.

Author Contributions: Conceptualization, D.A. and J.C.K.C.; methodology, Z.A., D.A. and J.C.K.C.; software, D.A.; formal analysis, Z.A. and J.C.K.C.; investigation, Z.A. and D.A.; resources, J.C.K.C.; writing—original draft preparation, Z.A. and J.C.K.C.; writing—review and editing, D.A., J.C.K.C. and Z.A.; visualization, J.C.K.C.; supervision, J.C.K.C.; funding acquisition, J.C.K.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the United States National Science Foundation, grant number 2017333.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Iowa State University, protocol 15-609. Approval date: 15 July 2020.

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: The original data presented in this study are openly available in Open Science Framework (OSF) at https://osf.io/8mwpn/ (last updated on 7 March 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

IRP	Interpolated retrieval practice
FTE	Forward testing effect
PI	Proactive interference
EEG	Electroencephalogram
ANOVA	Analysis of variance

References

Adhani, L. K., & Remijn, G. B. (2023, December 7–8). *A survey on external distractions affecting students' study performance during online learning*. 3rd International Conference on Community Engagement and Education for Sustainable Development (pp. 178–184), Grand Rohan Jogja Yogyakarta, Indonesia. [CrossRef]

Ahn, D., & Chan, J. C. K. (2022). Does testing enhance new learning because it insulates against proactive interference? *Memory & Cognition*, *50*, 1664–1682. [CrossRef]

Ahn, D., & Chan, J. C. K. (2024). Does testing potentiate new learning because it enables learners to use better strategies? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *50*(3), 435–457. [CrossRef]

Aivaz, K. A., & Teodorescu, D. (2022). College students' distractions from learning caused by multitasking in online vs. face-to-face classes: A case study at a public university in Romania. *International Journal of Environmental Research and Public Health*, *19*(18), 11188. [CrossRef] [PubMed]

Bjork, E. L., & Storm, B. C. (2011). Retrieval experience as a modifier of future encoding: Another test effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1113–1124. [CrossRef]

Boustani, S., Owens, C., Don, H. J., Yang, C., & Shanks, D. R. (2023). Evaluating the conceptual strategy change account of test-potentiated new learning in list recall. *Journal of Memory and Language*, *130*, 104412. [CrossRef]

Brame, C. J. (2016). Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE—Life Sciences Education*, *15*(4), 1–6. [CrossRef]

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*(1), 12–21. [CrossRef]

Chan, J. C. K., Ahn, D., Szpunar, K. K., Assadipour, Z., & Gill, H. (2025). Bridging the lab-field gap: Multi-site evidence that in-lecture quizzes improve online learning for university and community college students. *Communications Psychology*, *3*(1), 54. [CrossRef]

Chan, J. C. K., Manley, K. D., & Ahn, D. (2020). Does retrieval potentiate new learning when retrieval stops but new learning continues? *Journal of Memory and Language*, *115*, 104150. [CrossRef]

Chan, J. C. K., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018a). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective. *Journal of Memory and Language*, *102*, 83–96. [CrossRef]

Chan, J. C. K., & McDermott, K. B. (2007a). The effects of frontal lobe functioning and age on veridical and false recall. *Psychonomic Bulletin & Review*, *14*(4), 606–611. [CrossRef]

Chan, J. C. K., & McDermott, K. B. (2007b). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 431–437. [CrossRef] [PubMed]

Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018b). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin*, *144*(11), 1111–1146. [CrossRef]

Conrad, C., & Newman, A. (2021). Measuring mind wandering during online lectures assessed with EEG. *Frontiers in Human Neuroscience*, *15*, 697532. [CrossRef]

Davis, S. D., & Chan, J. C. K. (2015). Studying on borrowed time: How does testing impair new learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1741–1754. [CrossRef] [PubMed]

Davis, S. D., & Chan, J. C. K. (2023). Effortful tests and repeated metacognitive judgments enhance future learning. *Educational Psychology Review*, *35*(3), 86. [CrossRef]

Davis, S. D., Chan, J. C. K., & Wilford, M. M. (2017). The dark side of interpolated testing: Frequent switching between retrieval and encoding impairs new learning. *Journal of Applied Research in Memory and Cognition*, *6*(4), 434–441. [CrossRef]

Delaney, P. F., Sahakyan, L., Kelley, C. M., & Zimmerman, C. A. (2010). Remembering to forget: The amnesic effect of daydreaming. *Psychological Science*, *21*(7), 1036–1042. [CrossRef]

deWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition*, *32*(6), 945–955. [CrossRef]

Divis, K. M., & Benjamin, A. S. (2014). Retrieval speeds context fluctuation: Why semantic generation enhances later learning but hinders prior learning. *Memory & Cognition*, *42*(7), 1049–1062. [CrossRef]

Francis, M. K., Wormington, S. V., & Hulleman, C. (2019). The costs of online learning: Examining differences in motivation and academic outcomes in online and face-to-face community college developmental mathematics courses. *Frontiers in Psychology*, *10*, 2054. [CrossRef] [PubMed]

Glisky, E. L., Polster, M. R., & Routhieaux, B. C. (1995). Double dissociation between item and source memory. *Neuropsychology*, *9*(2), 229–235. [CrossRef]

Hansch, A., Hillers, L., McConachie, K., Newman, C., Schildhauer, T., & Schmidt, P. (2015). Video and online learning: Critical reflections and findings from the field. HIIG Discussion Paper Series No. 2015-02. [CrossRef]

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 290–296. [CrossRef]

Hew, K. F. (2018). Unpacking the strategies of ten highly rated MOOCs: Implications for engaging students in large online courses. *Teachers College Record*, *120*(1), 010308. [CrossRef]

Hollister, B., Nair, P., Hill-Lindsay, S., & Chukoskie, L. (2022). Engagement in online learning: Student attitudes and behavior during COVID-19. *Frontiers in Education*, *7*, 851019. [CrossRef]

Hong, M. K., Polyn, S. M., & Fazio, L. K. (2019). Examining the episodic context account: Does retrieval practice enhance memory for context? *Cognitive Research: Principles and Implications*, *4*(1), 46. [CrossRef]

Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 112–127. [CrossRef]

JASP Team. (2024). *JASP* (Version 0.18.3) [Computer Software]. Available online: https://jasp-stats.org (accessed on 5 May 2025).

Jing, H. G., Szpunar, K. K., & Schacter, D. L. (2016). Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of Experimental Psychology: Applied*, *22*(3), 305–318. [CrossRef]

Johnstone, A. H., & Percival, F. (1976). Attention breaks in lectures. *Education in Chemistry*, *13*, 49–50.

Jonker, T. R., Seli, P., & MacLeod, C. M. (2013). Putting retrieval-induced forgetting in context: An inhibition-free, context-based account. *Psychological Review*, *120*(4), 852–872. [CrossRef] [PubMed]

Kliegl, O., & Bäuml, K.-H. T. (2021). When retrieval practice promotes new learning—The critical role of study material. *Journal of Memory and Language*, *120*, 104253. [CrossRef]

Kliegl, O., & Bäuml, K.-H. T. (2023). How retrieval practice and semantic generation affect subsequently studied material: An analysis of item-level effects. *Memory*, *31*(1), 127–136. [CrossRef] [PubMed]

Kriechbaum, V. M., & Bäuml, K.-H. T. (2024). Retrieval practice can promote new learning with both related and unrelated prose materials. *Journal of Applied Research in Memory and Cognition*, *13*(3), 319–328. [CrossRef]

Lee, H. S., & Ahn, D. (2018). Testing prepares students to learn better: The forward effect of testing in category learning. *Journal of Educational Psychology*, *110*(2), 203–217. [CrossRef]

Murphy, D. H., Little, J. L., & Bjork, E. L. (2023). The value of using tests in education as tools for learning—Not just for assessment. *Educational Psychology Review*, *35*(3), 89. [CrossRef]

Pan, S. C., Schmitt, A. G., Bjork, E. L., & Sana, F. (2020). Pretesting reduces mind wandering and enhances learning during online lectures. *Journal of Applied Research in Memory and Cognition*, *9*(4), 542–554. [CrossRef]

Pastötter, B., Bäuml, K.-H., & Hanslmayr, S. (2008). Oscillatory brain activity before and after an internal context change—Evidence for a reset of encoding processes. *NeuroImage*, *43*(1), 173–181. [CrossRef]

Pastötter, B., & Bäuml, K.-H. T. (2014). Retrieval practice enhances new learning: The forward effect of testing. *Frontiers in Psychology*, *5*, 286. [CrossRef]

Pastötter, B., Schicker, S., Niedernhuber, J., & Bäuml, K.-H. T. (2011). Retrieval during learning facilitates subsequent memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 287–297. [CrossRef]

Peterson, L., & Peterson, M. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*(3), 193–198. [CrossRef] [PubMed]

Riccio, D. C., Richardson, R., & Ebner, D. L. (1984). Memory retrieval deficits based upon altered contextual cues: A paradox. *Psychological Bulletin*, *96*(1), 152–165. [CrossRef]

Riccio, D. C., Richardson, R., & Ebner, D. L. (1999). The contextual change paradox is still unresolved: Comment on Bouton, Nelson, and Rosas (1999). *Psychological Bulletin*, *125*(2), 187–189. [CrossRef]

Risko, E. F., Anderson, N., Sarwal, A., Engelhardt, M., & Kingstone, A. (2012). Everyday attention: Variation in mind wandering and memory in a lecture. *Applied Cognitive Psychology*, *26*(2), 234–242. [CrossRef]

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. [CrossRef] [PubMed]

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. [CrossRef] [PubMed]

Sahakyan, L., Delaney, P. F., Foster, N. L., & Abushanab, B. (2013). List-method directed forgetting in cognitive and clinical research. *Psychology of Learning and Motivation*, *59*, 131–189. [CrossRef]

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1064–1072. [CrossRef]

Schacter, D. L., Benoit, R. G., & Szpunar, K. K. (2017). Episodic future thinking: Mechanisms and functions. *Current Opinion in Behavioral Sciences*, *17*, 41–50. [CrossRef]

Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2014). Who does what in a massive open online course? *Communications of the ACM*, *57*(4), 58–65. [CrossRef]

Shimizu, Y., & Jacoby, L. L. (2005). Similarity-guided depth of retrieval: Constraining at the front end. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Experimentale*, *59*(1), 17–21. [CrossRef]

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, *73*, 99–115. [CrossRef]

Szpunar, K. K., Chan, J. C. K., & McDermott, K. B. (2009). Contextual processing in episodic future thought. *Cerebral Cortex*, *19*(7), 1539–1548. [CrossRef] [PubMed]

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1392–1399. [CrossRef]

Szpunar, K. K., Moulton, S. T., & Schacter, D. L. (2013). Mind wandering and education: From the classroom to online learning. *Frontiers in Psychology*, *4*, 495. [CrossRef] [PubMed]

Szpunar, K. K., Watson, J. M., & McDermott, K. B. (2007). Neural substrates of envisioning the future. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(2), 642–647. [CrossRef]

Tulving, E., Kapur, S., Craik, F. I., Moscovitch, M., & Houle, S. (1994). Hemispheric encoding/retrieval asymmetry in episodic memory: Positron emission tomography findings. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(6), 2016–2020. [CrossRef]

Tulving, E., & Watkins, M. (1974). Negative transfer—Effects of testing one list on recall of another. *Journal of Verbal Learning and Verbal Behavior*, *13*(2), 181–193. [CrossRef]

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335. [CrossRef]

Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 1039–1048. [CrossRef]

Weinstein, Y., McDermott, K., Szpunar, K. K., Bäuml, K.-H., & Pastötter, B. (2015, November 30). *Not all retrieval during learning facilitates subsequent memory encoding* [Conference presentation]. Annual Meeting of the Psychonomic Society, Chicago, IL, USA.

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning & Verbal Behavior*, *2*(5–6), 440–445. [CrossRef]

Yang, C., Chew, S.-J., Sun, B., & Shanks, D. R. (2019). The forward effects of testing transfer to different domains of learning. *Journal of Educational Psychology*, *111*(5), 809–826. [CrossRef]

Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied*, *23*(3), 263–277. [CrossRef]

Yang, C., Potts, R., & Shanks, D. R. (2018). Enhancing learning and retrieval of new information: A review of the forward testing effect. *Npj Science of Learning*, *3*(1), 8. [CrossRef] [PubMed]

Yang, C., Zhao, W., Luo, L., Sun, B., Potts, R., & Shanks, D. R. (2022). Testing potential mechanisms underlying test-potentiated new learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(8), 1127–1143. [CrossRef] [PubMed]

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory and Cognition*, *38*(8), 995–1008. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

The Distributed Practice Effect on Classroom Learning: A Meta-Analytic Review of Applied Research

Rhys D. Mawson and Sean H. K. Kang *

Faculty of Education, The University of Melbourne, Parkville, VIC 3010, Australia; rhys.mawson@hotmail.com
* Correspondence: sean.kang@unimelb.edu.au

Abstract

There is extensive evidence that distributed practice produces superior learning to massed practice, predominantly from laboratory studies often featuring decontextualized learning. A systematic review of applied research was undertaken to assess the impact of distributed practice on classroom learning. Inclusion criteria were classroom studies with learning materials and timescales relevant to curriculum-based learning. The screening of over 3000 articles resulted in 22 reports containing 31 effect sizes (N > 3000). A meta-analysis found a moderate effect in favor of distributed over massed practice (d = 0.54, 95% CI [0.31, 0.77]). Although a comprehensive quantitative moderator analysis was not possible due to the number of studies, generally larger effect sizes were associated with studies that featured longer retention intervals, had learners at higher education levels, and had fewer re-exposures to the materials.

Keywords: distributed practice; spaced practice; spacing effect; meta-analysis; classroom learning

1. Introduction

Although the trajectory of recent educational practices has moved away from the memorization of content (Pappas, 2009), repeated exposure to and the retention of foundational information is key to processing new information (Sweller, 2010). This remains true even in the context of 21st century classrooms aiming to promote 21st century skills (Saavedra & Opfer, 2012), where engagement with rich and complex tasks requires the scaffolded and timely retention of key knowledge (van Merrienboer et al., 2003). Thankfully, decades of research highlight several teaching and learning strategies that support and enhance learning and retention (Donoghue & Hattie, 2021; Dunlosky et al., 2013).

1.1. The Case for Distributed Practice

With a conceptualization refined over almost 150 years (Ebbinghaus, 1885/1913), distributed practice presents itself as a high-utility learning strategy (Dunlosky et al., 2013). Contrasted with *massed* practice, in which study material is presented in a back-to-back fashion, *distributed* or 'spaced' practice involves spacing out the same duration of study time across multiple study sessions (Rohrer & Pashler, 2010). Distributed practice involves the initial learning of content followed by one or more study sessions separated by a set time (i.e., the interstudy interval).

However, distributed practice is more than just learning spread out over time. Indeed, although some studies operationalize spaced practice as 'stretched out' study time (e.g., 26 h over 4.5 days versus 3 days; Breckwoldt et al., 2016), this misses the critical factor

driving the benefits of spaced over massed practice. Specifically, theoretical explanations for the effect of distributed practice emphasize that it is the *repeated exposure to material* spaced over time that leads to efficient learning (Kang, 2016). For example, encoding variability theory suggests that information is likely to be encoded in different ways when repetitions are spaced (instead of massed), leading to multiple retrieval pathways and increased accessibility; according to the study-phase retrieval account, when an item is repeated after a gap in time, the learner is potentially reminded of its prior occurrence, prompting retrieval of the previous presentation, which enhances memory (Maddox, 2016). The deficient processing of massed repetitions is another theory—i.e., when information is repeated in quick succession, less attention is paid to the repetitions (Gerbier & Toppino, 2015). But this does not mean that more exposures automatically equal better retention. The nature of the gap between exposures also plays a role in the effect of distributed practice (Wiseheart et al., 2019). Indeed, evidence suggests that 'expanding' or increasingly long gaps between exposures have a positive effect on learning in some situations (Kang, 2016). In addition to spacing, the number of exposures is also important, with evidence that too many exposures can lead to fatigue that negatively impacts learning (Gurung & Burns, 2019). Although implementing the multiple exposures required to engage in 'true' distributed practice requires careful consideration, the consistently positive effect of distributed practice makes it a 'desirable difficulty' for both learners and educators (Bjork & Bjork, 2011).

Indeed, meta-analytic research emphasizes the benefit of distributed over massed practice on the retention of information (Cepeda et al., 2006; Donoghue & Hattie, 2021; Donovan & Radosevich, 1999; Janiszewski et al., 2003; Lee & Genovese, 1988). Moreover, the effect of this and other learning strategies are consistent across a range of characteristics of both the learners and learning involved, including age, learning domain, and ability level (Donoghue & Hattie, 2021). However, despite recent widespread research and communication regarding its positive impact (e.g., Dunlosky & Rawson, 2015; Kang, 2016; Rohrer & Pashler, 2010), distributed practice has a long history of being underutilized in educational settings (Dempster, 1988).

1.2. Bridging the Research–Practice Divide

The limited application of distributed practice likely stems from a disconnect between how learning is conceptualized in experimental settings and real-world classrooms (Ebersbach et al., 2022). Indeed, the 'traditional' laboratory-based study of distributed practice often involves very short retention (e.g., <10 min) and interstudy intervals (e.g., five seconds to one week) and is based predominantly on verbal learning tasks (see Cepeda et al., 2006; Janiszewski et al., 2003). Recent work to distill the relevance of this highly constrained research to educators does exist (Department of Education and Training, 2020; Pomerance et al., 2016). However, the highly specific evidence base for distributed practice means questions remain about the strategy's application to real-world classrooms (Ebersbach et al., 2022).

Applied research into distributed practice has no doubt taken place, but its effect is often conflated with that of laboratory-based studies. For example, Donoghue and Hattie's (2021) meta-analysis of learning strategies indicated a strong (d = 0.85) effect of distributed practice across studies including over 150,000 unique participants. While the authors reviewed the effect of this and other learning strategies across key variables, they did not distinguish between effect sizes drawn from classroom versus laboratory-based studies. As such, it is difficult to provide guidance on how to implement the strategy when there is no clear picture of the factors that influence its effect—positively and negatively—in everyday classrooms. As evidenced by similar research into other learning strategies (e.g., retrieval

practice; Agarwal et al., 2021), a specific focus on classroom-based research could more clearly identify the effect of distributed practice in real-world settings.

Moreover, research conducted in educational settings more effectively captures the factors that influence real-world learning outcomes (Berliner, 2002). One of these contexts relevant to distributed practice is the various timescales of real-world learning. For example, it is increasingly common for school-based curriculum to require that students "progressively apply the familiar to develop new ideas" across year levels (Australian Curriculum Assessment and Reporting Authority, n.d.). In addition to this longer-term accumulation of knowledge, students also need to learn information in order to tackle upcoming classroom tasks and homework, show their knowledge at the end of units of work, or pass future exams (Foot-Seymour et al., 2019). Similarly, real-world classroom learning involves a range of materials and methods. In today's classrooms, for example, students learn via approaches such as project-based learning (Lonergan et al., 2022), tackle 21st century skills like critical thinking (Foot-Seymour et al., 2019), and are often asked to go beyond surface-level learning (Hattie & Donoghue, 2016). Applied research into distributed practice helps bridge the research–practice gap by actively contending with the contexts of classroom learning.

Importantly, an up-to-date review would capture the recent resurgence of applied learning strategy research that is missing from previous reviews. Most notably, Donoghue and Hattie's (2021) meta-analysis is based predominantly on the articles explored in Dunlosky et al.'s (2013) previous review of the learning strategy literature. Although comprehensive, the articles current at the time of Dunlosky et al.'s (2013) review are now more than 10 years old. Critically, there has been a growth of applied research on learning strategies in recent decades, likely a reflection of educators' need to maximize limited classroom time (Agarwal et al., 2021; Kang, 2016). As such, the numerous articles published in the preceding 10 years likely add valuable information relevant to the application of distributed practice to classroom learning (e.g., Barzagar Nazari & Ebersbach, 2019a; Barzagar Nazari & Ebersbach, 2019b; Foot-Seymour et al., 2019; Greving & Richter, 2021; Lyle et al., 2019; Svihla et al., 2017). An updated review specifically focused on distributed practice research in applied settings is therefore both timely and feasible.

1.3. The Present Review

While the effect of distributed practice on learning is strong in laboratory-based studies, the effectiveness of this strategy on real-world classroom learning is less clear-cut. As such, a focus on applied research is required to paint a more accurate picture of the effectiveness of distributed practice on classroom learning. Moreover, a review focused on applied research would capture key aspects of classroom learning that may moderate the effect of distributed practice. These variables are critical to the successful and effective implementation of such a strategy.

The present review has two main objectives. First, this review aims to quantify the effect of distributed practice on primary, secondary, and university-level students' ability to learn content relevant to their everyday classrooms. In addition, this review aims to explore the characteristics of classroom-based learning that may moderate the effect of distributed practice.

2. Method

PRISMA guidelines (see Page et al., 2021) were followed in this systematic review. Minor deviations (see, e.g., Section 2.6) are noted and explained in the sections outlined below.

2.1. Eligibility Criteria

Inclusion criteria were developed to focus this review on the application of distributed practice in mainstream educational settings. These criteria are summarized in Table 1 below. At the research level, the gray literature was not eligible for inclusion except for theses and dissertations. This concession was made because current information services systems mean theses and dissertations are more readily available forms of the literature than they have been previously (Schopfel & Rasuli, 2018). Moreover, doctoral dissertations in the field of educational psychology have a long history of being applied in focus (e.g., Ysseldyke & Pickholtz, 1975). Theses and dissertations therefore reflect an increasingly accessible source of information relevant to the aim of the present study.

Table 1. Summary of eligibility criteria for the systematic review.

	Inclusion	**Exclusion**
Population	Study participants enrolled in primary, secondary, or tertiary education	Participants drawn exclusively from special populations (e.g., autistic, intellectual disability, etc.)
Intervention	Distributed or spaced practice with interstudy intervals of at least 1 day	Interstudy intervals of less than 1 day
Comparison	Massed practice conditions with time spent studying equal to that of spaced practice conditions	Comparison to other learning strategies
Outcomes	Measures of retention (e.g., percentage correctly recalled) or transfer of information learnt (e.g., percentage correctly applied). Where retention was measured more than once, data were extracted from the longest retention interval measured in both the comparison and intervention conditions	Qualitative outcomes (e.g., teacher and/or student perceptions of effectiveness)
Study design and context	Studies conducted in the classroom using classroom-relevant learning materials	Studies using learning materials with no clear connections to specific syllabi or curriculum

In terms of study samples, the widespread emphasis on inclusive practices in education ensures that a wide range of learners—with and without disabilities—are participating and succeeding in mainstream classes (Anderson & Boyle, 2015). While research solely focused on special populations (e.g., Robinson, 1979) is important for a more nuanced understanding of the impact of this strategy and other learning strategies, it was excluded on the basis that it less effectively represents the broad types of classrooms that exist in mainstream learning environments.

For the intervention, a one-day interstudy interval was chosen as the smallest interval applicable to real-world learning. Specifically, everyday classrooms tend to cluster the learning of content, focusing on introducing new material in one class and then practicing it in the next lesson a day or so later (Svihla et al., 2017). A one-day interstudy interval is therefore relevant to the way that classrooms currently operate. Moreover, recent large-scale research suggests that optimal learning outcomes occur when the interstudy interval is approximately 10–20% of the required retention interval (Cepeda et al., 2008, 2009). While this indicates that a retention period of 7 days provides the most effective use of practice distributed over a day, retention periods both more and less than 7 days were included (i.e.,

no exclusions were placed on retention interval). This decision was motivated by the fact that the goal or purpose of learning defines the length of time needed to retain it, a factor which likely varies considerably across study levels, subject areas, and learning tasks.

Related to the study design criteria, the concept of 'classroom-relevant learning materials' was deliberately kept broad. This was due to the idea that relevance is defined by the learning environment, both in terms of the learners and learning domain. For example, pre-defining that the learning needed to be assessed formally (e.g., through exams) would have excluded primary school-aged learners whose learning is often more informally and qualitatively assessed. Similarly, a mathematics classroom in a primary school likely has access to very different resources and approaches to a graduate-level university course in accounting, in terms of how learning is presented, practiced, and assessed; methods for the distribution of practice and assessment of learning did not factor into exclusion and inclusion of studies, so long as they did not deviate substantially from what those learners might be expected to experience in their classroom. In general, studies were included if they took place in the classroom and involved content that would have been learnt by students in that class regardless of whether researchers had intervened.

2.2. Search Strategy

Articles were drawn from the Education Resources Information Centre (ERIC; ProQuest), PsycINFO (OVID), A+ Education (Informit), and ProQuest Central (ProQuest) databases. These databases were searched from their inception to June 2023. Common search terms were used for each database, including variants of *spaced*, *distributed*, *practice*, and *learning*. Changes were made to syntax as required by individual database guides. The search syntax used for each individual database is presented in Appendix A.

Citation searches were also used to supplement articles drawn from databases. Initially, reference lists included in recent meta-analyses and systematic reviews of the relevant literature were scanned for classroom-based studies meeting criteria (i.e., Donoghue & Hattie, 2021; Ebersbach et al., 2022). Following the initial screening and full-text review, a reference list search of the articles meeting inclusion criteria was also undertaken.

2.3. Selection Process

To monitor the inclusion and exclusion process, EndNote v20.6 (The EndNote Team, 2023) was used. One researcher (RM) reviewed the title and abstract of all studies to determine eligibility for full-text review. To monitor fidelity to the eligibility criteria, 10% of articles eligible for full-text review were also reviewed by the supervising researcher (SK).

To facilitate this, the citations for all full-text, non-duplicated reviews written in English (n = 194) were copied to Microsoft Excel (Microsoft Corporation, 2023). Excel's 'random number between' function was used to generate 19 numbers between 2 and 195, corresponding with the row numbers of citations. The 19 associated titles, abstracts, and full-text articles were then independently reviewed by SK. Disagreements were resolved through discussion and consensus.

2.4. Data Items and Collection

The data extraction process was undertaken by one researcher (RM). This process involved the extraction of a range of data from the included literature, including study elements (e.g., year of publication), information about participants (e.g., education level), and key features of both the learning (e.g., learning domain or subject) and learning outcome (i.e., M and SD values for retention of information across conditions).

2.5. Synthesis Methods

To assess the effect of distributed versus massed practice, a Standardized Mean Difference (SMD) was calculated for every comparison in each study. Data relevant to mean retention at the longest retention interval were tabulated for the distributed and massed practice conditions in each study within a report. The mean retention of the massed condition was subtracted from that of the spaced condition and divided by a pooled standard deviation to create an effect size (i.e., Cohen's *d*). Using this approach, positive values represent an effect in favor of distributed practice.

In all cases, reported means and standard deviations were used to calculate SMD values using an online effect size calculator (Wilson, n.d.). This process took place regardless of whether effect sizes were reported in individual studies, as researchers have a tendency to calculate effect sizes in multiple and sometimes conflicting ways (Lakens, 2013). This ensured consistency in the way that effect sizes were calculated. The random effect meta-analysis of the SMD values and their standard errors was conducted in R (v4.3.1; R Core Team, 2023) using the *metafor* package (v4.2-0; Viechtbauer, 2010). The decision to assume random effects was made given the significant differences in approaches to teaching and learning across countries, states, and individual schools. Output relevant to heterogeneity (e.g., I^2 statistics) and its significance was also sought. The R code is included in the Supplementary Materials.

2.6. Bias and Certainty Assessment

Applied studies have an inherent focus on ecological validity. Hence, although high-quality applied experiments include elements of randomized-controlled trials, the two designs have distinct aims. As such, an explicit risk of bias assessment was deemed irrelevant. In terms of reporting bias, a funnel plot of effect sizes against standard errors for included studies was created for visual inspection. No certainty assessment was undertaken.

3. Results

3.1. Screening

As highlighted in the flowchart (adapted from Page et al., 2021) in Figure 1 below, almost 90% of studies assigned to a full-text review were excluded. In addition to the application of the inclusion criteria above, studies were also excluded based on other features. For example, the greatest number of studies were excluded based on their study design. Here, studies were often decontextualized in comparison to real-world classroom learning, either because they were based in a lab (e.g., in a 'simulated' classroom; Kapler et al., 2015), used materials irrelevant to the class (e.g., primary school students studying vocabulary words from a postgraduate vocabulary reasoning exam; Sobel et al., 2011), or employed a combination of both (e.g., Suzuki & Dekeyser, 2017). In contrast, included studies involved learning tasks where the content was directly related to the specific course being studied (e.g., the statement of cash flows in an accounting course; McNellis, 2015) or was directly relevant to the general age and stage of the learner in more general classrooms (e.g., relevant spelling words for learners in a primary-school classroom; Petersen-Brown et al., 2023). In addition, a common theme was studies claiming to apply distributed practice when they were better described as examples of block scheduling. For example, several studies involving second-language word learning had students review groups of words either in one long session or across several shorter sessions (Namaziandost et al., 2019, 2020, 2023). However, these studies did not include repeated exposure to any one word. This represents a common misconception of distributed practice.

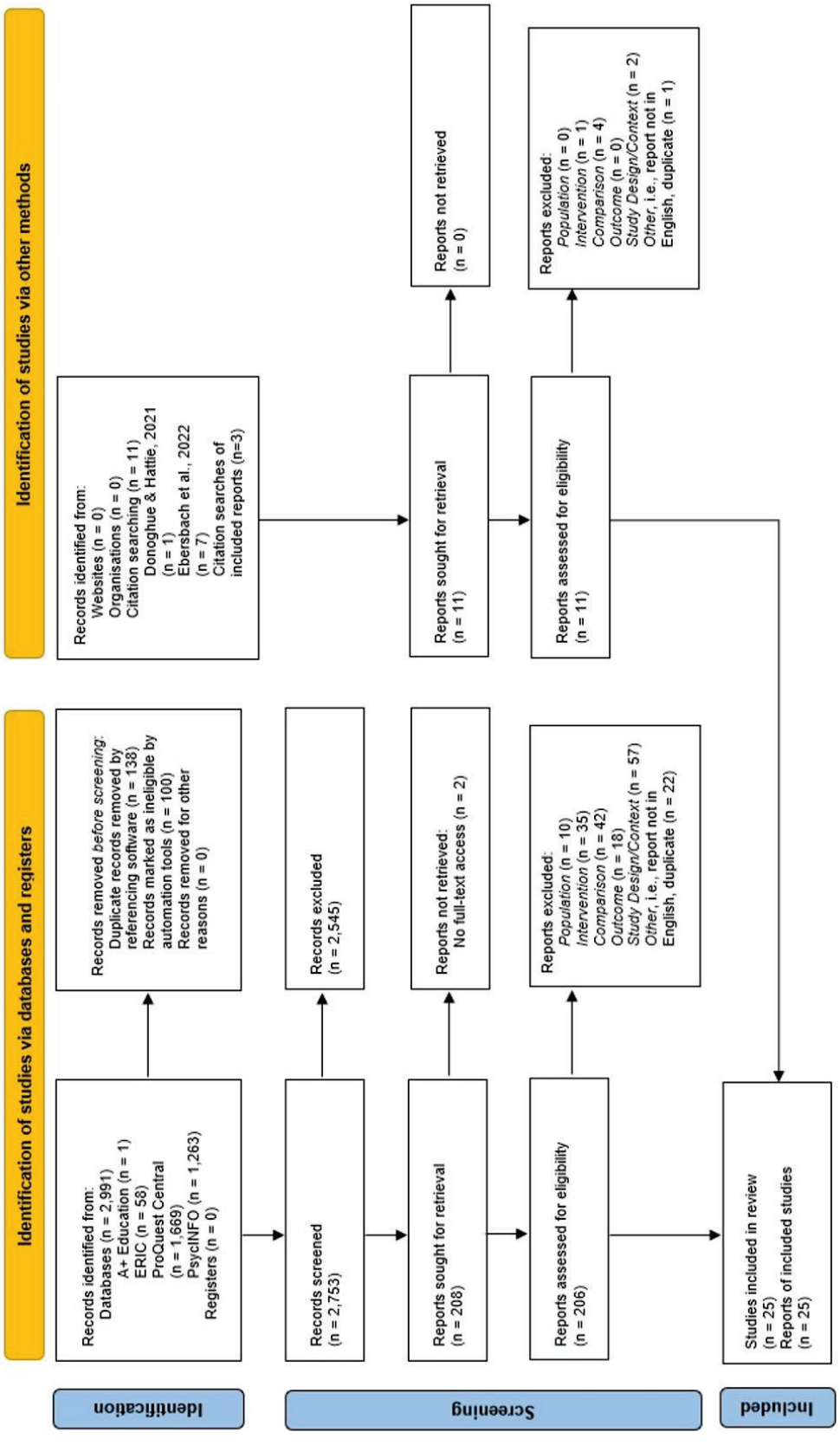


Figure 1. Flowchart of the screening process.

Another area of contention was the use of comparison groups that did not involve massed practice, made more challenging by the open definition of this condition in the inclusion criteria. For example, while some studies used easily excluded 'business-as-usual' conditions that involved neither massed nor distributed practice (e.g., Benson et al., 2022), others compared distributed practice to distributed 'retrieval'. For example, Gjerde et al. (2022) had some students read through solutions to physics problems (review-only), while others were given problems to solve themselves (retrieval practice or 'spaced testing'; see Agarwal et al., 2021). The lack of a massed practice comparison, and not the use of retrieval practice itself, led to the exclusion of this study. Similarly, distributed and massed practice conditions were included in Lindsey et al.'s (2014) study, but results from both were combined to compare against interleaved practice (for a definition and review, see Dunlosky et al., 2013). Overall, studies were excluded when distributed and massed practice conditions were not identifiable and clearly distinguishable.

3.2. Included Studies

3.2.1. Included in Review

In total, 25 studies met all inclusion criteria. The full table of data extracted from the 25 studies is included under Supporting Materials.

3.2.2. Included in Analysis

Despite 25 studies meeting criteria for review, not all studies were eligible for inclusion in a quantitative synthesis. For three reports (Ebersbach & Barzagar Nazari, 2020; Grote, 1995; and Barzagar Nazari & Ebersbach, 2018), outcome data could not be extracted or retrieved via contact with the authors. This resulted in 22 studies eligible for analysis.

Additional factors were also considered before conducting the quantitative analysis. This included the presence of unspecified or unequal retention periods, predominantly in which the timing between final exposure and the retention test was shorter for the distributed practice condition. In addition, non-independent effects existed in several included studies. While this was overcome in two studies by focusing only on results for retention tests and not those assessing students' application of critical thinking skills (Foot-Seymour & Wiseheart, 2022; Foot-Seymour et al., 2019), non-independent effects remained in other studies. In those cases, additional standardized mean difference (SMD) values were created by combining the M and SD values across measures within each condition (i.e., massed and distributed) to create a single, independent effect. These combined values were calculated using the method outlined in the Cochrane Handbook for Systematic Reviews of Interventions (i.e., Chapter 6; Higgins et al., 2023) with the aid of an online tool (https://www.statstodo.com/CombineMeansSDs.php, accessed on 20 September 2023).

Effect sizes related to the above factors are clearly distinguished in the data extraction table (see Supplementary Materials). The potential impact of including or excluding those effects was explored through a preliminary analysis.

3.3. Synthesis

3.3.1. Preliminary Analysis

Four preliminary meta-analyses were conducted (see Table 2 below). Each analysis employed a restricted Maximum Likelihood model under random-effect assumptions.

All analyses indicated a positive effect size, indicating overall results in favor of distributed over massed practice. In terms of differences, the qualitative comparison of Analyses 1 and 2 compared with 3 and 4 indicates that any overestimation due to unequal retention intervals in favor of the distributed practice condition is relatively small. In addition, replacing non-independent effect sizes with a single, average effect size for each

relevant study resulted in a slight increase in the overall combined effect size (i.e., Analyses 2 and 4). Despite this, overall estimates based on the inclusion of non-independent effects (i.e., Analyses 1 and 3) are likely more accurate given that they are based on a greater number of effect sizes.

Table 2. Summary of meta-analyses conducted.

Analysis	Potential Data Inaccuracies		Studies	Effect Sizes	Effect Size [95% CI]	I^2 (%)
	Unequal Retention Intervals	Non-Independent Effect Sizes				
1	Included	Included All	22	31	0.54 [0.31, 0.77]	92.13
2	Included	Averaged	22	25	0.63 [0.36, 0.89]	94.93
3	Excluded	Included All	16	23	0.51 [0.31, 0.72]	83.88
4	Excluded	Averaged	16	19	0.57 [0.36, 0.78]	84.68

Acknowledging the above points, Analysis 1 appeared to maximize the amount of information for analysis while minimizing the inaccuracy introduced by averaging nonindependent effect sizes. As such, Analysis 1 is described in detail below.

3.3.2. Main Analysis

As confirmed visually in the forest plot in Figure 2 below, standardized mean difference (SMD) values across included studies showed an overall significant effect of distributed over massed practice (d = 0.54, 95% CI [0.31, 0.77], z = 4.631, $p < 0.001$). This indicates that the mean retention for learners in distributed practice conditions was, on average, over half a standard deviation higher than for those in massed practice conditions.

Of note, Egger's test (z = 2.719, p = 0.007) confirms the significant asymmetry in the distribution of effects visible in Figure 3 below. However, as indicated by the highlighted areas of statistical significance, many studies were non-significant. This suggests that the asymmetry is unlikely due to publication bias. Instead, the asymmetry likely represents a 'small study bias', where methodological approaches inherent in small studies lead to stronger effects with greater variance (Peters et al., 2008; Sterne et al., 2011). The large (d = 3.11; n = 120) effect in the study by Yazdani and Zebrowski (2006) is a notable exception.

Notably, heterogeneity can also influence funnel plot asymmetry (Sterne et al., 2011). In the main analysis, the significant I^2 value shown in Table 2 suggests that more than 92% of the variance between studies was due to systematic influence and not chance ($Q(30)$ = 230.33, $p < 0.001$). A moderator analysis via meta-regression was considered as one method for exploring this systematic influence.

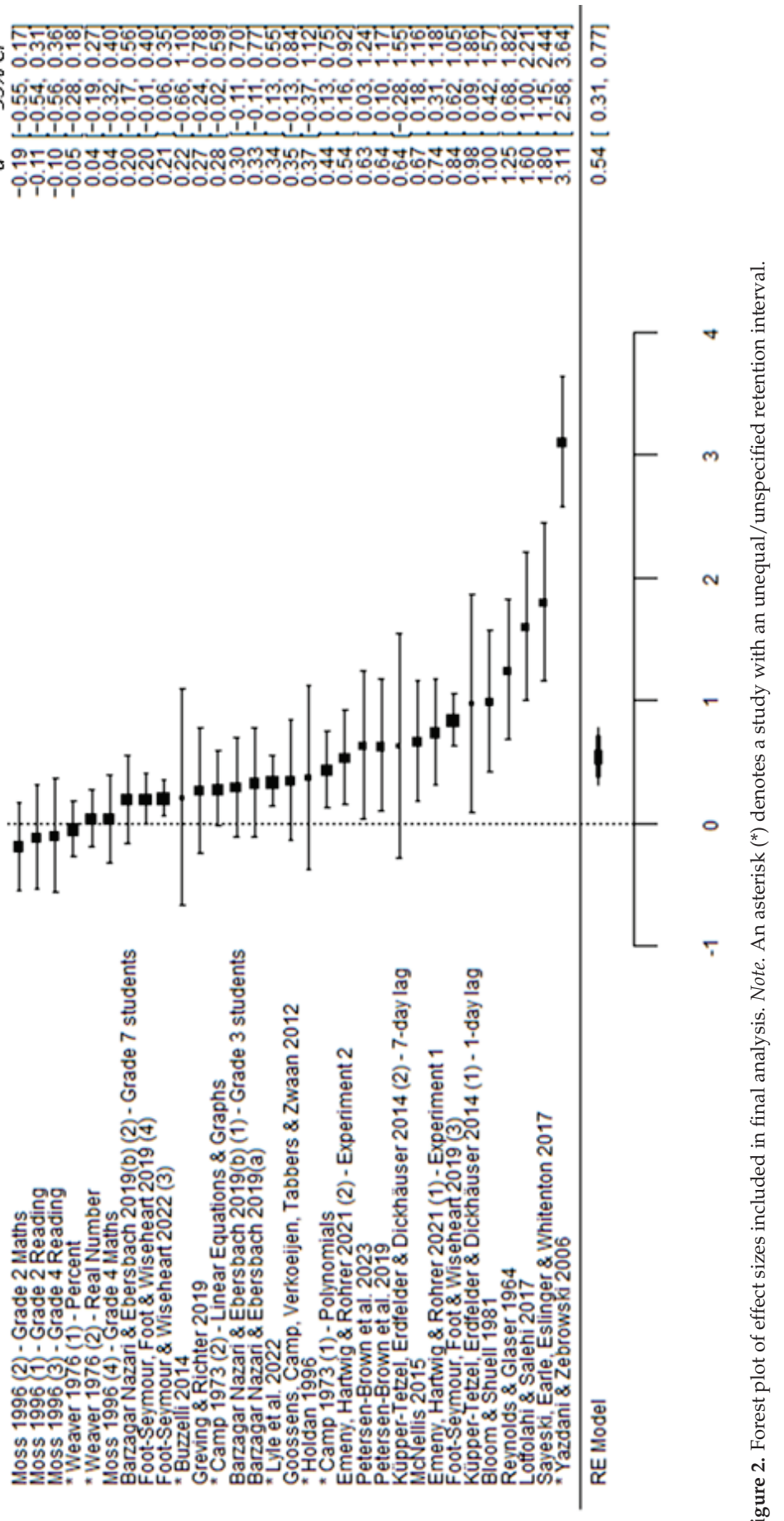


Figure 2. Forest plot of effect sizes included in final analysis. *Note.* An asterisk (*) denotes a study with an unequal/unspecified retention interval.

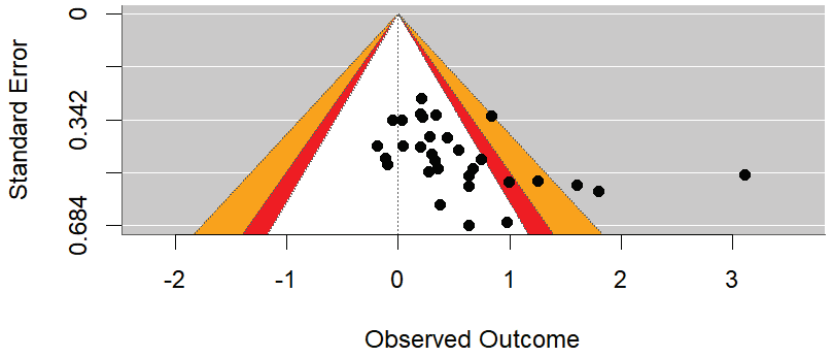


Figure 3. Contour-enhanced funnel plot of effect sizes from studies included in analysis (k = 22). *Note*. Orange = $p < 0.01$, red = $p < 0.05$, and white = $p < 0.10$ (i.e., statistically non-significant).

3.3.3. Moderator Analysis

Several variables related to the timescales of and materials for learning were chosen a priori and extracted to facilitate a moderator analysis. However, guidelines recommend that 10 studies are required for every variable entered into the meta-analysis (i.e., two variables as appropriate for k = 22 studies); moreover, this is only true when there is an even distribution of covariates within a variable or characteristics (Deeks et al., 2023). Of the timescale variables, the data extraction table in the Supplementary Materials shows that none contained an appropriate amount of variability and/or were widely and accurately reported. For example, most studies used a fixed interstudy interval, while the amount of total study time was not reported in many studies. In addition, the retention interval could not be used as a moderator, given that several of the included studies made use of unequal or unspecified retention intervals. Of the learning material variables, 'Learning Domain' was the only factor widely reported and with good variability in levels.

When Analysis 1 was re-run with 'Learning Domain' as a moderating variable, results revealed a non-significant difference in the effect size at each level ($Q(8) = 7.56$, $p = 0.478$). This indicates that the estimated effect sizes within the nine learning domains were not significantly different from one another. This is evidenced by the overlap in estimated effect sizes within each learning domain, as shown in Table 3 below.

Table 3. Results of the random-effect meta-regression with 'Learning Domain' as a moderator.

Learning Domain	Effect Size [95% CI]
Accounting	0.67 [0.18, 1.16]
Critical Thinking	0.41 [0.00, 0.82]
History	0.22 [0.00, 0.44]
Language	0.91 [0.44, 1.3]
Math	0.46 [0.09, 0.84]
Reading	−0.11 [−0.42, 0.21]
Science	0.75 [−0.21, 1.71]
Spelling	0.63 [0.03, 1.24]
Teacher Education	1.80 [1.15, 2.44]

Note. Meta-regression model uses the 'Accounting' level of the variable as the reference level.

Moreover, the presence of significant levels of residual heterogeneity ($Q(22) = 180.85$, $p < 0.001$; $I^2 = 92.84\%$) indicates that this variable does not adequately explain the heterogeneity found in the main analysis.

The small amount of extracted data, as well as difficulties in extracting the required information from all studies, precluded additional moderator analyses. However, when grouped under key variables, qualitative differences in the size and direction of effect

sizes were apparent. Effect sizes in the included literature were therefore grouped and reviewed based on relevant extracted variables (as per Agarwal et al., 2021). This approach is taken to illustrate possible trends in effects that, although beyond the scope of this study, require statistical exploration to understand whether *qualitative* differences are supported by *quantitative* analysis.

3.4. Key Variables

3.4.1. Learning Domain

As seen in Figure 4, consistently large effect sizes came from language learning studies. However, most effect sizes in this domain were drawn from classrooms focused on second-language learning. Indeed, only the single study by Goossens et al. (2012) suggested that the positive effect of distributed practice may generalize to vocabulary learning more generally. In comparison, the math learning domain had a larger number of studies, less variability in effects, and generally positive effect sizes for distributed over massed practice. Given that the included studies in this domain captured a range of math topics, distributed practice appears generally effective in the math domain.

3.4.2. Retention Interval

Figure 5 highlights that the effect of distributed over massed practice tended to be larger at longer retention intervals. However, the present review did not find studies with intervals longer than 42 days, indicating a lack of information about the effects of distributed practice over very long-term scales. Of note, most studies with either unequal or unspecified retention intervals were still associated with positive albeit small effect sizes. However, the majority created conditions that favored those in the distributed practice condition, with final distributed practice sessions often taking place closer to the final test than the massed practice session. The fact that this 'advantage' was not associated with stronger effects in favor of distributed practice suggests the presence of other factors in these studies (see Section 4).

3.4.3. Interstudy Interval

Fixed intervals of seven days between re-exposures tended to produce the most consistently positive and significant effects for distributed over massed practice (see Figure 6). In addition, the effect sizes for expanding interstudy intervals tended to be smaller and, in one case, negative. However, the effect sizes for expanding intervals included here may not generalize, as most are based on the same study procedure.

3.4.4. Education Level

Figure 7 shows that studies in primary school settings produced the largest number of negative effects (i.e., favoring massed over distributed practice). Moreover, studies at this education level had the highest proportion of non-significant effect sizes. In contrast, secondary school-level studies produced more consistently positive and significant effects. Interestingly, despite previous lab-based distributed practice research predominantly employing university-level participants, the present review captured only a small number of applied studies in university-level classrooms.

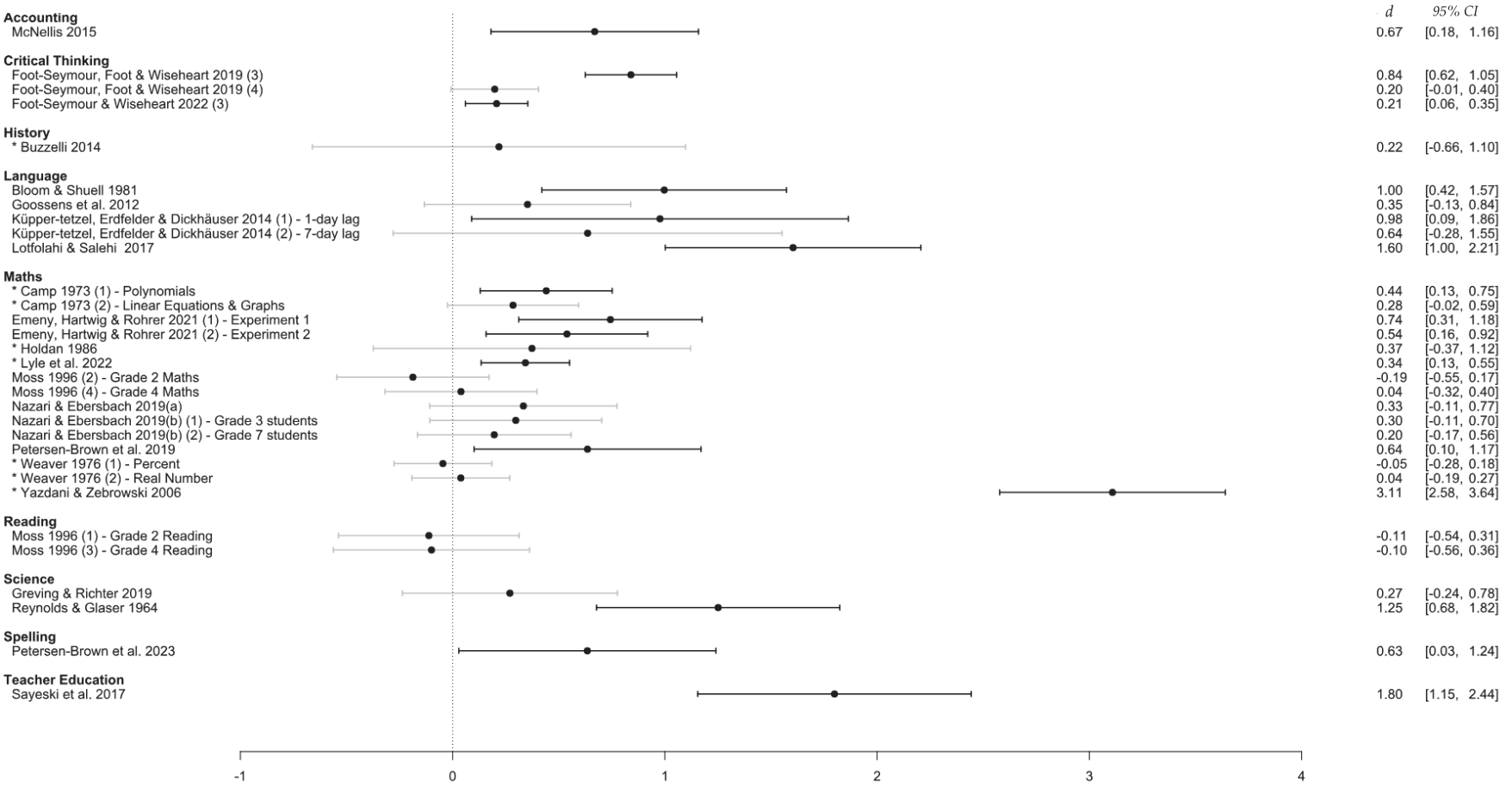


Figure 4. Effect sizes grouped by learning domain. *Note.* This figure shows the distribution of effect sizes when grouped by the area of learning (e.g., subject, course, or topic) that the content falls within. Grayed-out confidence intervals indicate those that cross the line of no difference (i.e., statistically non-significant). An asterisk (*) denotes a study with an unequal/unspecified retention interval.

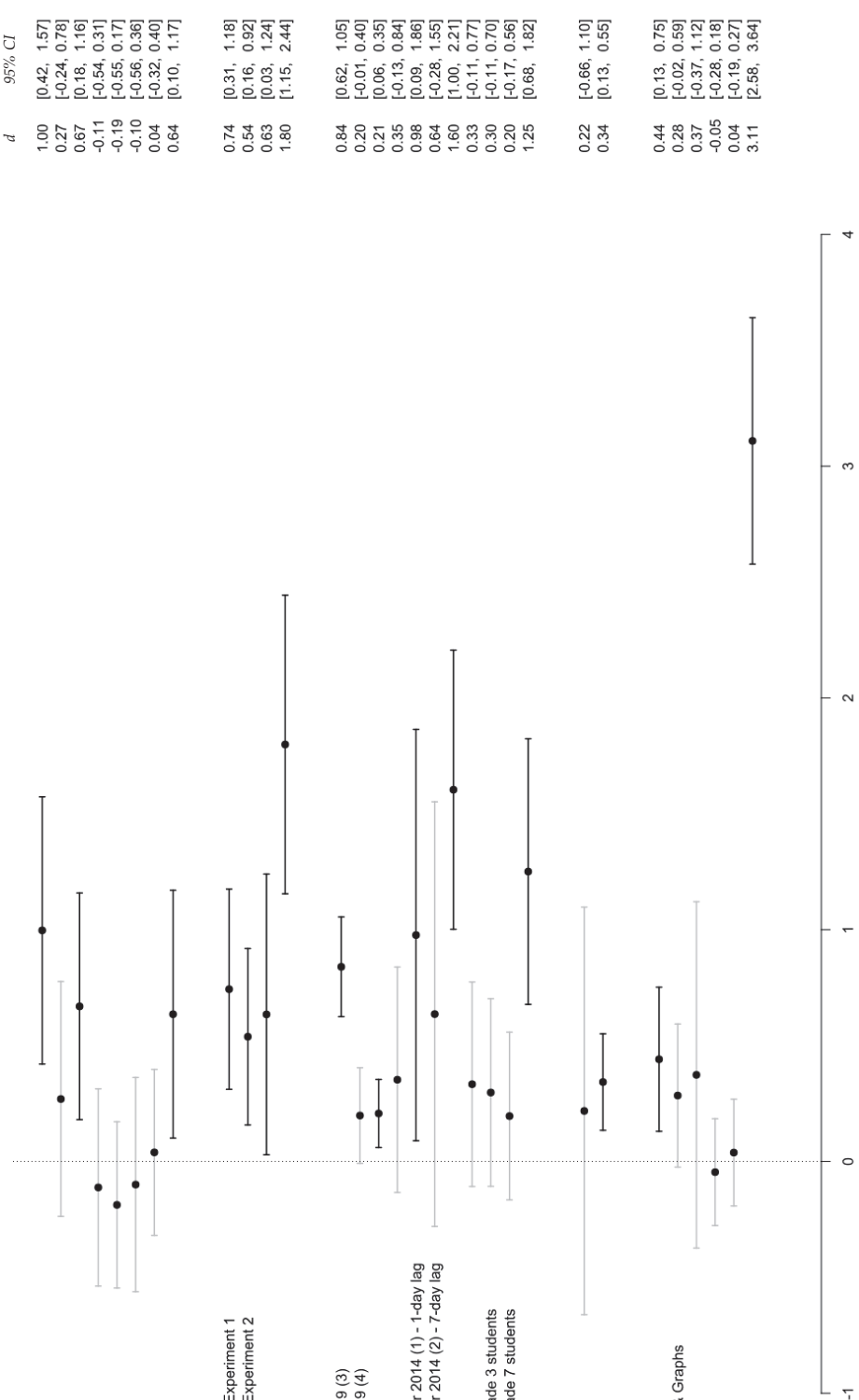


Figure 5. Effect sizes grouped by retention interval category. *Note.* This figure shows the distribution of effect sizes when grouped based on the length of time that participants have to retain the information, from the last study session to the assessment of learning outcomes. An asterisk (*) denotes a study with an unequal/unspecified retention interval.

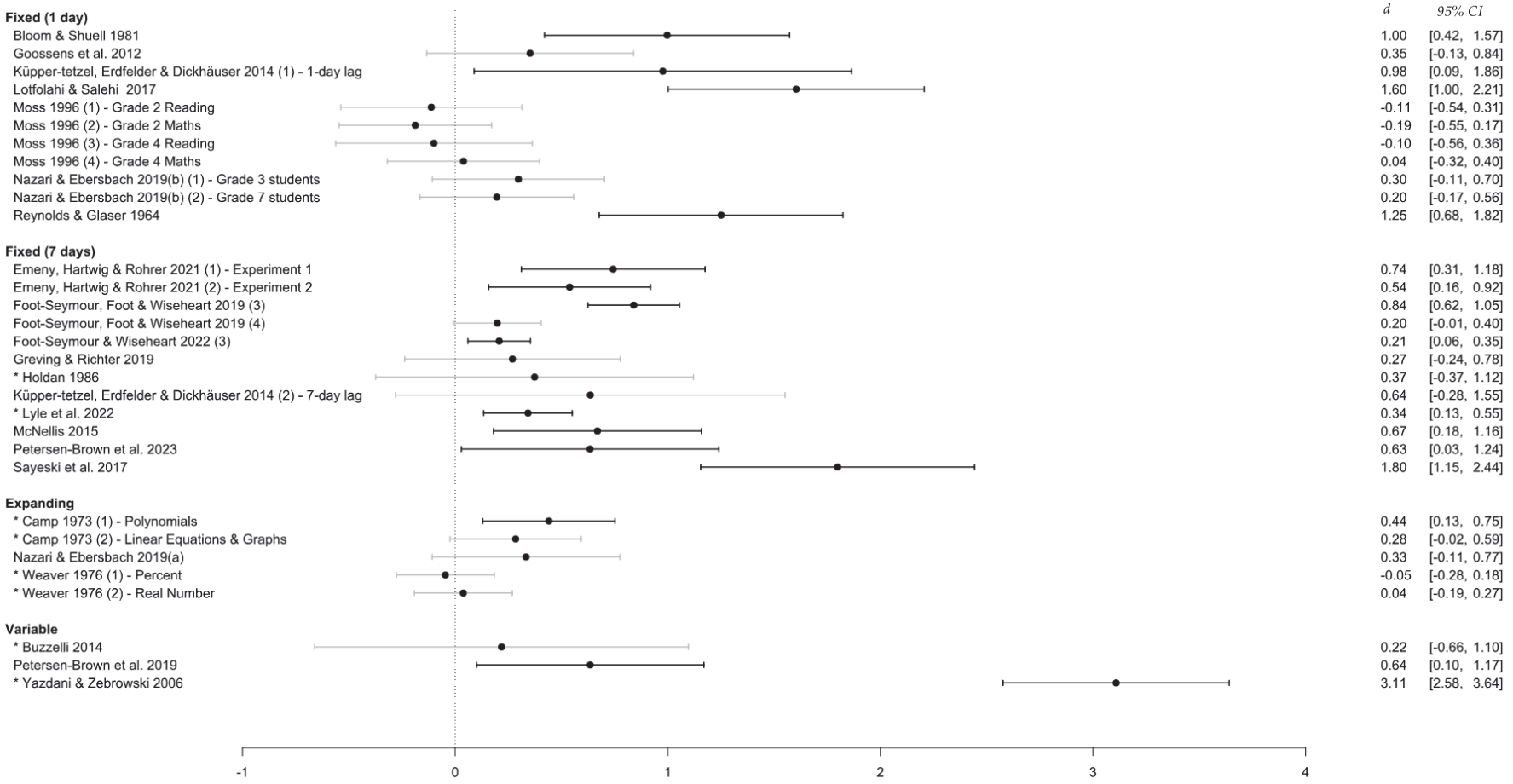


Figure 6. Effect sizes grouped by interstudy interval category. *Note.* This figure shows the distribution of effect sizes based on the length (1 versus 7 days) and nature (fixed versus expanding [i.e., increasing time between each re-exposure]) of the gap between study sessions. An asterisk (*) denotes a study with an unequal/unspecified retention interval.

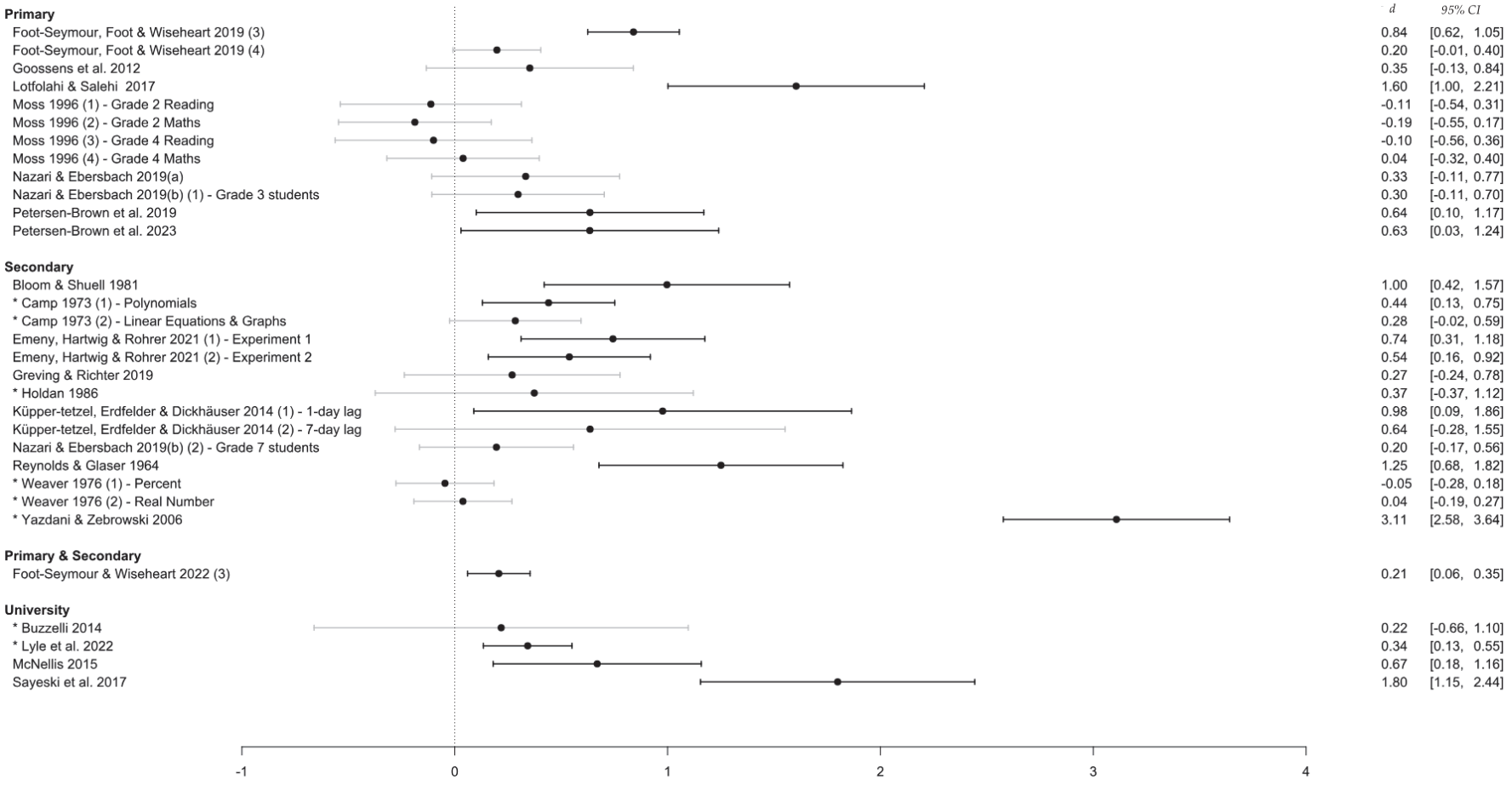


Figure 7. Effect sizes grouped by education level. *Note.* This figure shows the distribution of effect sizes based on whether participants are classified as primary-, secondary-, or university-level students by this study's authors based on the cut-off ages and/or year levels used by the country in which this study took place. An asterisk (*) denotes a study with an unequal/unspecified retention interval.

3.4.5. Number of Re-Exposures

Generally speaking, a smaller number of practice sessions resulted in a stronger effect in favor of distributed over massed practice. Indeed, Figure 8 shows that most studies with three exposures after the initial learning were associated with significant and moderate-to-large effect sizes. Of note, four of the five effect sizes drawn from distributed practice over more than three sessions came from one study. Indeed, the number of re-exposures used in Moss (1996) was considerably larger than much of the included literature (i.e., nine re-exposures to each of the reading and math learning material). While it was not possible to ascertain the exact number of re-exposures due to a lack of reporting, most of the studies in the 'unclear' level of this variable also described a procedure that included a high number of exposures (i.e., up to 5). Given that these effect sizes were also smaller than those of studies with fewer exposures, it is likely that these studies provide further evidence for the trend of 'less is more' when it comes to practice sessions.

3.5. Key Themes

In addition to the above variables, a broader look at the included literature revealed several key themes. These themes represent areas of the design of the included studies, which may provide helpful information for implementing distributed practice in real-world classrooms.

3.5.1. The 'Practice' in Distributed Practice

While all studies involved practice via repeated exposure to the learning material, the form of that 'exposure' differed greatly across the literature. For example, in studies utilizing distributed practice for language or vocabulary learning (e.g., Bloom & Shuell, 1981), repeated exposure was simply a re-presentation of the same words. In contrast, the studies involving math learning tended to present practice problems that, although based on the same content, were not the exact same problems with the exact same numbers. The study by Lyle et al. (2022), for example, described the different practice problems as "algorithmic variants" (p. 1803). Although repeatedly exposed to the same general idea, learners in this and similar studies rarely—if ever—saw the exact same thing in each exposure. In real-world classrooms, the nature of repeated exposure appears to depend on the content being learnt.

The *process* of these repeated exposures also differed across studies. For example, some studies focused solely on repeated exposure (e.g., re-reading the same passage; Greving & Richter, 2019), others involved learners working on problems associated with the initially learnt content (e.g., practice math problems; Weaver, 1976), while others still asked learners to review and practice content through the use of complex exercises (e.g., the Cover–Copy–Compare process for learning spelling words; Petersen-Brown et al., 2023). Moreover, the studies aiming to improve students' critical thinking (i.e., Foot-Seymour et al., 2019; Foot-Seymour & Wiseheart, 2022) had the additional complexity of explicitly asking students to transfer or apply information to completely new situations. These variations on the process of practice are all highly relevant to the learning process in real-world classrooms.

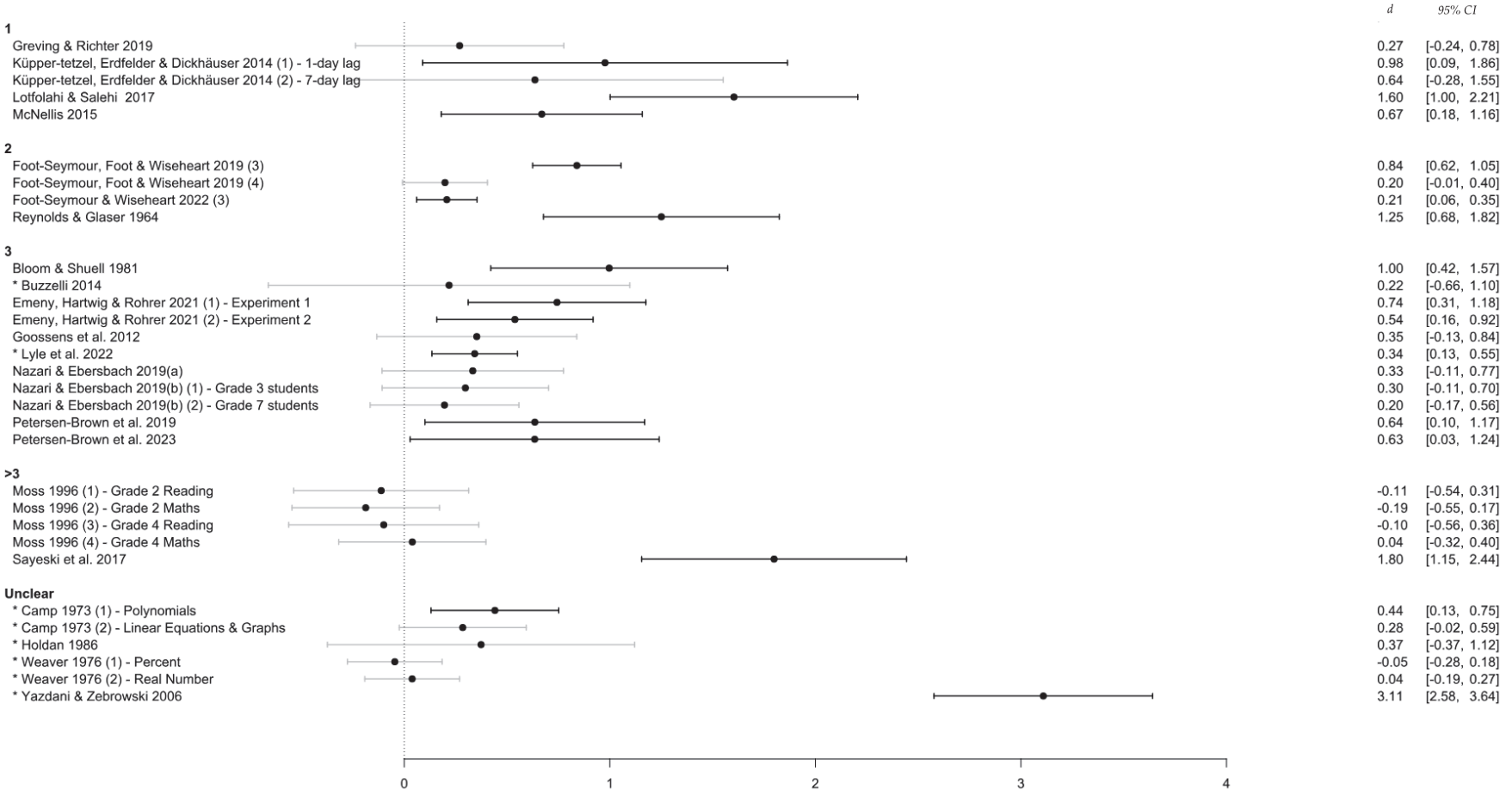


Figure 8. Effect sizes grouped by number of re-exposures. *Note.* This figure shows the distribution of effect sizes based on whether participants received 1, 2, 3, or more than three exposures to the content in the distributed practice condition. An asterisk (*) denotes a study with an unequal/unspecified retention interval.

3.5.2. Feedback in Practice and the Role of Technology

Moreover, the nature of how the learning material was practiced also influenced the extent to which learners could receive feedback on their learning across exposures. In particular, the provision of feedback depended on the extent to which repeated exposures asked students to remember or simply review the information. For example, Buzzelli's (2014) use of tweets via Twitter focused on giving students the information presented in earlier lectures in a different, additional form. As such, each 'practice' session or exposure did not require action from learners other than reading. In contrast, the study by Petersen-Brown et al. (2019) involved learners giving definitions for exposed words, and, when they provided incorrect definitions, they were given feedback and then asked to correctly recall the definition again later. As such, these learners also received more exposure than those who correctly recalled the definitions initially. Learning to criterion (i.e., where an item correctly recalled a set number of times is removed from further review) has been criticized as an unnecessary confounding factor in distributed practice research (Cepeda et al., 2006); it is one that appears to have made its way into applied studies. In general, feedback was more common in studies that involved active engagement from learners rather than just passive review.

In addition, the ability for learners to receive feedback was heavily influenced by whether technology formed a part of the practice process. For example, while practice problems were a common theme in studies of math learning, these problems were presented in different ways. The use of an online math-practice platform in the study by Lyle et al. (2022) enabled learners to receive feedback on practice problems after completing the relevant quizzes. In contrast, participants in studies by Camp (1973) and Holdan (1986) did not receive feedback on completed textbook questions. In studies involving non-math learning, the use of digital flashcards in Sayeski et al.'s (2017) study also enabled real-time feedback. Moreover, the feedback was presented by a peer, something also employed in other included studies but without technology (i.e., Lotfolahi & Salehi, 2017). Where the feedback comes from adds further complexity to the question of how feedback influences the success of distributed practice specifically and learning more generally.

The provision of feedback across distributed practice sessions also highlights the interrelationship between spaced *practice* and spaced *testing*. This has been highlighted previously as a widely used strategy in today's classrooms (Agarwal et al., 2021). Given that the combination of these two related approaches to repeated exposure can influence learning (Gurung & Burns, 2019), clearly distinguishing between the effects of these two strategies is key to the success of their respective implementations in classrooms. Indeed, a key differentiating factor may be the extent to which the repeated exposure is facilitated by the teacher (i.e., testing) or the learner themselves (i.e., studying).

3.5.3. Distributed Practice as a Teaching or Learning Strategy?

The literature varied widely in the extent to which distributed practice was viewed as a teaching versus learning strategy. None of the included studies gave learners the option to practice materials in a distributed fashion without requiring it, a necessary consequence of experimental studies in which procedures are pre-determined and adherence is expected. Instead, studies differed in the extent that teachers drove the practice sessions. For example, approaches included sessions in which learners practiced on their own (e.g., Camp, 1973), those in which peers practiced with each other without a teacher (e.g., Lotfolahi & Salehi, 2017; Sayeski et al., 2017), studies involving teachers working alongside individual students (e.g., Petersen-Brown et al., 2019), and those in which teachers directly 'taught' the content in each exposure (e.g., McNellis, 2015). However, the positive effects across this range

of implementation approaches likely reflect the versatility of distributed practice in real-world classrooms.

4. Discussion

4.1. The Overall Effect

Compared to massed practice, distributed practice had a moderate effect (d = 0.54) on classroom-based learning of curriculum-relevant materials. This provides clear evidence in favor of using distributed practice to support everyday classroom learning. However, results of the present study indicate a somewhat smaller overall effect in comparison to previous reviews (d = 0.85; Donoghue & Hattie, 2021). The fact that the Donoghue and Hattie (2021) meta-analysis combined the effects of distributed practice in experimental and applied settings is a likely factor in this difference.

Differences in effect sizes across such designs have previously been attributed to differences in the complexity of the learning that takes place in them (Ebersbach et al., 2022). Indeed, previous meta-analyses of distributed practice highlight that its effect is often stronger for less complex tasks, where complexity is based on the requirements of and decisions to be made within the task (Donovan & Radosevich, 1999). This fits with the findings of the present review.

Specifically, a key component of complexity in classroom learning is what learners need to do with the knowledge gained. While the decision was made in the present review to focus solely on the retention of information, several included studies employed learning and practice that went beyond just restating previous information. For example, most studies in the included literature involved math learning and used distributed practice conditions in which learners completed practice questions. However, each question was not simply a restatement of previous questions and tended to include slight variations, such as different values or contexts. As effect sizes drawn from math-learning studies tended to be smaller and more frequently non-significant in the present review, it is possible that the added complexity of needing to transfer learning to relatively novel situations meant that such studies captured something more complex than 'surface level' learning (Hattie & Donoghue, 2016). The availability of clear and easily applied definitions for levels of learning is therefore important to future reviews of applied distributed practice research.

However, the complexity of materials is unlikely to be the only factor relevant for educators seeking to effectively implement distributed practice. Indeed, a qualitative synthesis of the included studies highlighted differences in effect sizes across a range key variables.

4.2. Timing as a Factor

As expected, based on previous laboratory-based research, the present study indicated that the interstudy interval remains a key factor in the size of the effect of distributed practice over massed practice. For example, studies with a 7-day interstudy interval were consistently statistically significant and often associated with moderate-to-large effect sizes when compared to massed practice. However, in real-world classrooms, most teaching and learning tend to be condensed into distinct units that are covered intensely over a short number of days, rather than being covered in a single longer session as is done in massed practice conditions (Svihla et al., 2017). To better understand the effect of timing on distributed practice, future research could use alternative comparison conditions (e.g., 1-day interstudy intervals) that align more closely to classroom learning.

Similarly, the higher frequency of larger effects for studies with longer intervals found in this study broadly reflects the pattern seen in laboratory studies. However, such research also emphasizes that the effect of retention intervals on the distributed practice advantage

does not continuously increase as the retention interval increases (Cepeda et al., 2008). Also, there is evidence from lab experiments with just one re-exposure that the optimal interstudy interval depends on the retention interval (Cepeda et al., 2009). However, the relationship between interstudy and retention intervals was not something that could be assessed with the included data, given the somewhat restricted nature of retention intervals employed (i.e., <42 days) and the varying numbers of re-exposures. Ongoing research in applied settings would benefit from a focus on this relationship.

Finally, studies with unequal or unspecified retention intervals require individual consideration. While retention interval decisions in these studies were somewhat constrained by efforts to fit the experiment into the existing school calendar, they nevertheless created conditions that appeared to favor distributed practice conditions (i.e., shorter retention intervals). Despite this apparent advantage, studies with unequal retention intervals were associated with smaller effect sizes, even compared to those with which they shared other methodological aspects (e.g., same learning domain). Influential factors here may have been the common use of expanding interstudy intervals or the higher number of re-exposures used in these studies. The negative impact of these factors fits with evidence that, in scheduling too many exposures to content, learning can be negatively impacted through increasing students' fatigue from constantly reviewing the content (Gurung & Burns, 2019).

While highlighting the complexities of both planning for and implementing effectively timed distributed practice, the present study nevertheless captures a range of time scales more relevant to real-world classrooms. Critically, results suggest that distributed practice remains broadly effective across this range.

4.3. Learning Domain and Level

The magnitude—but not direction—of effect sizes favoring distributed practice appears to differ across learning domain (i.e., subject area). However, unlike previous reviews (i.e., Donoghue & Hattie, 2021), the meta-regression in the present did not confirm the statistical significance of this apparent difference. This may have been caused by differences in the variability of learning domains captured in the included studies relative to previous research. For example, studies involving learning in English or Reading made up 35.1% of the included studies in Donoghue and Hattie's (2021) review but only 6% in the present review. The larger proportion in the former may reflect that the included laboratory-based studies involving language-based learning (e.g., verbal paired associates) were considered most closely aligned to those learning domains. Generally speaking, areas beyond language and math learning appear underrepresented in the applied literature, and, except for the studies applying distributed practice to critical thinking (Foot-Seymour et al., 2019; Foot-Seymour & Wiseheart, 2022), distributed practice is yet to be widely applied to 21st-century skills learning. These represent critical areas for future study.

In terms of education level, studies set in secondary schools were more often associated with stronger positive effects of distributed practice on learning than those in primary school settings. Although smaller in magnitude, this difference in effects is reflected in the spread of data in previous meta-analyses, where the effect of distributed practice was generally smaller for primary school learners (mean Cohen's $d = 0.57$) than for those in secondary schools (mean Cohen's $d = 0.70$; Donoghue & Hattie, 2021). However, although viewing effect sizes through the lens of educational levels confirms stronger positive effects for older children, descriptive statistics for age were not widely reported.

As such, the present review could not untangle whether positive effects are due to some element of learning environments involving older learners or a variable associated with age more generally. For example, most secondary schools tend to compartmentalize learning

into separate units within isolated subjects, albeit this is a trend that more progressive learning approaches such as project-based learning attempt to reshape (Lonergan et al., 2022). It is therefore possible that, in using more cross-curricular approaches to learning, primary students engage in more complex learning that is less amenable to distributed practice. Similarly, the nature of how classes are timetabled at secondary schools versus primary schools differs significantly, particularly the way in which secondary schools focus on more discreetly and inflexibly blocked learning times or how primary schools tend to keep the same groups of students together across classes. These aspects of classrooms at each education level could impact both the study of distributed practice in applied research and the effects found within. Isolating the effects of age from educational level, as well as a closer look at the nature of the learning environment at each level, is likely to be a fruitful avenue for future research.

4.4. Limitations

Although providing evidence for the application of distributed practice in everyday classrooms, the present review has some limitations. For example, despite the decision to use narrow search criteria, our preliminary analysis indicated significant levels of heterogeneity in the included literature. This may have been driven by the fact that some elements of the inclusion criteria were not well defined. Specifically, distributed practice was defined based on timing rather than on the process and content of the practice itself. As discussed in the Results Section, this led to comparisons between disparate studies that varied in terms of how content was re-exposed and what learners needed to do with it. Further refinement of a definition for the 'practice' in distributed practice may decrease heterogeneity in future reviews.

In addition, the methods for both the calculation and interpretation of effect sizes in this study may have influenced our results. Indeed, the methods used to create effect sizes may have introduced inaccuracy into the measurement of the distributed practice effect. Specifically, the calculation of standardized mean effect sizes was conducted using extractable data, rather than relying on the effect sizes presented in individual studies. While this decision was made to avoid the various ways that effect sizes are calculated in the literature generally (Lakens, 2013) and in learning strategy research specifically (see Agarwal et al., 2021), such an approach may mean that the critical information influencing each study author's statistical decision making was lost. Also, the interpretation of effect sizes across variables in this study was based on *qualitative* rather than *quantitative* analysis. This decision was heavily influenced by the lack of statistical power in the collated data. While this study provides an important initial step in understanding differences in effects within variables, larger sample sizes are required for a formal moderator analysis.

In a handful of the studies that focused on mathematics, practice assignments in the distributed practice condition—but not the massed practice condition—featured a mixture of different topics (Camp, 1973; Holdan, 1986; Weaver, 1976; possibly Yazdani & Zebrowski, 2006). In other words, in a few cases, distributed practice was conflated with interleaved practice, which itself has been shown to benefit mathematics problem solving (e.g., Taylor & Rohrer, 2010), and, thus, it is unclear to what extent the effects observed in these studies are due to spacing vs. interleaving. Having said that, it is worth noting that there is evidence that the interleaved practice effect in mathematics is caused by distributed practice (Foster et al., 2019).

Finally, the present study is unable to make specific comments on the effect of such a strategy on special populations (e.g., autistic learners). It can be assumed that students with learning and developmental disorders would be included in representative samples of inclusive classrooms. However, data relevant to sample representativeness were not

reported, and, hence, the assumption cannot be guaranteed. Also, effect sizes are calculated from group averages and do not tell us whether or which subset(s) of the sample may have benefited to a different degree from distributed practice. Future research could consider participant samples more closely or focus on special populations specifically.

4.5. Implications

4.5.1. Educational Implications

The generally positive effect of distributed practice—across a wide range of approaches to that practice—provides clear evidence that re-exposure to material in any form across separate sessions is likely to improve learning. While results of the present study highlight the benefit of this strategy for math learning and in secondary school settings in particular, findings more generally suggest that distributed practice can—and should—be broadly applied. As such, educators across curriculum areas and within schools that use a range of teaching methods are encouraged to explore their use.

Importantly, this research emphasizes that the interplay between timing variables, as well as their respective and combined influence on the effect of distributed practice, represents a complex set of considerations for educators. Indeed, one of the challenges in the included literature was the impact of needing to fit the distributed practice around predefined tasks such as end-of-unit assessments or semester exams. Far from discouraging the application of distributed practice in real-world classrooms, the present study highlights the importance of embedding distributed practice during the planning phase of creating individual units and broader subjects. In particular, teachers are encouraged to think about when the learning will be used (e.g., for a test at the end of a four-week unit) to help them make decisions about spacing out the associated practice.

4.5.2. Research Implications

Despite being a clearly stated aim of the present study, the quality of the available research—in both its focus and reporting—hampered the effort to create practical guidance for educators looking to implement this strategy. For one, the present study illustrates a disconcerting level of unclearly reported or missing data across key variables. This impacts the capacity of researchers to offer defensible implementation suggestions even on variables widely considered to impact the effect of this strategy. Alongside more consistent reporting of those variables, it is critical that researchers actively question whether these are the variables of most interest to educators. For example, the present study provides preliminary—but quantitatively unsubstantiated—information on what elements of learning are being distributed and how, as well as the role that technology and feedback play in its spacing and retrieval. A widespread focus on these factors will enable researchers to provide advice to educators that is more relevant and easily implemented, across curriculum areas and levels of study.

5. Conclusions

Distributed practice plays a positive and significant role in classroom learning outcomes. Undoubtedly, its implementation requires consideration of the nature of the learning and the timescales over which it needs to take place. Nevertheless, embedding distributed practice into the planning of units and subjects enables teachers and students across educational levels and curriculum areas to reap the benefits of this strategy.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/bs15060771/s1: full tables of extracted data and R code. References (Emeny et al., 2021; Küpper-tetzel et al., 2014; Reynolds & Glaser, 1964) are cited in the supplementary materials.

Author Contributions: Conceptualization, S.H.K.K.; methodology, R.D.M. and S.H.K.K.; software, R.D.M.; formal analysis, R.D.M.; investigation, R.D.M. and S.H.K.K.; data curation, R.D.M.; writing—original draft preparation, R.D.M.; writing—review and editing, S.H.K.K.; visualization, R.D.M.; supervision, S.H.K.K.; project administration, R.D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data supporting reported results will be made available by the authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Search syntax by database.

Database	Search Syntax	Search Fields
A+ Education (Informit)	(“spacing effect” OR “distribut* practi?e”~1 OR “spac* practi?e”~1 OR “mass* practi?e”~1 OR “distribut* learn*”~1 OR “spac* learn*”~1 OR “mass* learn*”~1) AND (class* OR stud* OR learn OR appli*)	Default
ERIC	(“spacing effect” OR “(distribut* OR spac* OR mass*) N/1 practi#e” OR “(distribut* OR spac* OR mass*) N/1 learn*”) AND (class* OR stud* OR appli*)	Default (All Fields+ text)
PsycINFO (Ovid)	((spacing effect or ((distribut* or spac* or mass*) adj practi#e) or ((distribut* or spac* or mass*) adj learn*)) and (class* or stud* or appli*))	Default (Title, Abstract, Heading Word, Table of Contents, Key Concepts, Original Title, Tests & Measures, Mesh Word)
ProQuest Central (ProQuest)	(“spacing effect” OR “(distribut* OR spac* OR mass*) N/1 practi#e” OR “(distribut* OR spac* OR mass*) N/1 learn*”) AND (class* OR stud* OR appli*)	Default (All Fields+ text)

References

Note: Studies included in the meta-analytic review are denoted by *

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*, *33*(4), 1409–1453. [CrossRef]

Anderson, J., & Boyle, C. (2015). Inclusive education in Australia: Rhetoric, reality and the road ahead. *Support for Learning*, *30*(1), 4–22. [CrossRef]

Australian Curriculum Assessment and Reporting Authority. (n.d.). *F-10 curriculum: Mathematics—Key ideas*. Available online: https://www.australiancurriculum.edu.au/f-10-curriculum/mathematics/key-ideas/ (accessed on 1 October 2023).

* Barzagar Nazari, K., & Ebersbach, M. (2018). Distributed practice: Rarely realized in self-regulated mathematical learning. *Frontiers in Psychology*, *9*, 2170. [CrossRef]

* Barzagar Nazari, K., & Ebersbach, M. (2019a). Distributed practice in mathematics: Recommendable especially for students on a medium performance level? *Trends in Neuroscience and Education*, *17*, 100122. [CrossRef]

* Barzagar Nazari, K., & Ebersbach, M. (2019b). Distributing mathematical practice of third and seventh graders: Applicability of the spacing effect in the classroom. *Applied Cognitive Psychology*, *33*(2), 288–298. [CrossRef]

Benson, W. L., Dunning, J. P., & Barber, D. (2022). Using distributed practice to improve students' attitudes and performance in statistics. *Teaching of Psychology*, *49*(1), 64–70. [CrossRef]

Berliner, D. C. (2002). Comment: Educational research: The hardest science of all. *Educational Researcher*, *31*(8), 18–20. [CrossRef]

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 55–64). Worth Publishers.

* Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *The Journal of Educational Research*, *74*(4), 245–248. [CrossRef]

Breckwoldt, J., Ludwig, J. R., Plener, J., Schroder, T., Gruber, H., Peters, H., Breckwoldt, J., Ludwig, J. R., Plener, J., Schroder, T., Gruber, H., & Peters, H. (2016). Differences in procedural knowledge after a "spaced" and a "massed" version of an intensive course in emergency medicine, investigating a very short spacing interval. *BMC Medical Education*, *16*, 249. [CrossRef]

* Buzzelli, A. A. (2014). *Twitter in the classroom: Determining the effectiveness of utilizing a microblog for distributed practice in concept learning* [Doctoral dissertation, Robert Morris University].

* Camp, J. S. (1973). The effects of distributed practice upon learning and retention in introductory algebra. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *34*(5-A), 2455–2456.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*(4), 236–246. [CrossRef] [PubMed]

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. [CrossRef]

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–1102. [CrossRef] [PubMed]

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2023). Chapter 10: Analysing data and undertaking meta-analyses. In J. P. T. Higgins, & J. Thomas (Eds.), *Cochrane handbook for systematic reviews of intervention (version 6.4)*. Cochrane. Available online: https://training.cochrane.org/handbook/current (accessed on 2 June 2023).

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*(8), 627–634. [CrossRef]

Department of Education and Training. (2020). *High impact teaching strategies: Excellence in teaching and learning*. Available online: https://www.education.vic.gov.au/school/teachers/teachingresources/practice/improve/Pages/hits.aspx (accessed on 12 April 2023).

Donoghue, G. M., & Hattie, J. A. C. (2021). A meta-analysis of ten learning techniques. *Frontiers in Education*, *6*, 581216. [CrossRef]

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795–805. [CrossRef]

Dunlosky, J., & Rawson, K. A. (2015). Practice tests, spaced practice, and successive relearning: Tips for classroom use and for guiding students' learning. *Scholarship of Teaching and Learning in Psychology*, *1*(1), 72–78. [CrossRef]

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. [CrossRef]

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*. Teachers College, Columbia University.

* Ebersbach, M., & Barzagar Nazari, K. (2020). Implementing distributed practice in statistics courses: Benefits for retention and transfer. *Journal of Applied Research in Memory and Cognition*, *9*(4), 532–541. [CrossRef]

Ebersbach, M., Lachner, A., Scheiter, K., & Richter, T. (2022). Using spacing to promote lasting learning in educational contexts. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *54*(4), 151–163. [CrossRef]

* Emeny, W. G., Hartwig, M. K., & Rohrer, D. (2021). Spaced mathematics practice improves test scores and reduces overconfidence. *Applied Cognitive Psychology*, *35*(4), 1082–1089. [CrossRef]

* Foot-Seymour, V., Foot, J., & Wiseheart, M. (2019). Judging credibility: Can spaced lessons help students think more critically online? *Applied Cognitive Psychology*, *33*(6), 1032–1043. [CrossRef]

* Foot-Seymour, V., & Wiseheart, M. (2022). Judging the credibility of websites: An effectiveness trial of the spacing effect in the elementary classroom. *Cognitive Research: Principles and Implications*, *7*, 5. [CrossRef]

Foster, N. L., Mueller, M. L., Was, C., Rawson, K. A., & Dunlosky, J. (2019). Why does interleaving improve math learning? The contributions of discriminative contrast and distributed practice. *Memory & Cognition*, *47*, 1088–1101.

Gerbier, E., & Toppino, T. C. (2015). The effect of distributed practice: Neuroscience, cognition, and education. *Trends in Neuroscience and Education*, *4*(3), 49–59. [CrossRef]

Gjerde, V., Vegard Havre, P., Holst, B., & Kolstø, S. D. (2022). Problem solving in basic physics: Effective self-explanations based on four elements with support from retrieval practice. *Physical Review Physics Education Research*, *18*(1), 010136. [CrossRef]

* Goossens, N. A., Camp, G., Verkoeijen, P. P., Tabbers, H. K., & Zwaan, R. A. (2012). Spreading the words: A spacing effect in vocabulary learning. *Journal of Cognitive Psychology*, *24*(8), 965–971. [CrossRef]

* Greving, C. E., & Richter, T. (2019). Distributed learning in the classroom: Effects of rereading schedules depend on time of test. *Frontiers in Psychology*, *9*, 2517. [CrossRef]

Greving, C. E., & Richter, T. (2021). Beyond the distributed practice effect: Is distributed learning also effective for learning with non-repeated text materials? *Frontiers in Psychology*, *12*, 685245. [CrossRef]

* Grote, M. G. (1995). Distributed versus massed practice in high school physics. *School Science and Mathematics*, *95*(2), 97–101. [CrossRef]

Gurung, R. A. R., & Burns, K. (2019). Putting evidence-based claims to the test: A multi-site classroom study of retrieval practice and spaced practice. *Applied Cognitive Psychology*, *33*(5), 732–743. [CrossRef]

Hattie, J. A. C., & Donoghue, G. M. (2016). Learning strategies: A synthesis and conceptual model. *npj Science of Learning*, *1*, 16013. [CrossRef] [PubMed]

Higgins, J. P. T., Li, T., & Deeks, J. J. (2023). Choosing effect measures and computing estimates of effect. In J. P. T. Higgins, & J. Thomas (Eds.), *Cochrane handbook for systematic reviews of intervention (version 6.4)*. Cochrane. Available online: https://training.cochrane.org/handbook/current (accessed on 2 June 2023).

* Holdan, E. G. (1986). A comparison of the effects of traditional, exploratory, distributed, and a combination of distributed and exploratory practice on initial learning, transfer, and retention of verbal problem types in first-year algebra. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *46*(9-A), 2542. Available online: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc2&NEWS=N&AN=1986-55259-001 (accessed on 16 May 2023).

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research*, *30*(1), 138–149. [CrossRef]

Kang, S. H. K. (2016). Spaced repetition promotes efficient and effective learning. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 12–19. [CrossRef]

Kapler, I. V., Weston, T., & Wiseheart, M. (2015). Spacing in a simulated undergraduate classroom: Long-term benefits for factual and higher-level learning. *Learning and Instruction*, *36*, 38–45. [CrossRef]

* Küpper-tetzel, C. E., Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science*, *42*(3), 373–388. [CrossRef]

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. [CrossRef]

Lee, T. D., & Genovese, E. D. (1988). Distribution of practice in motor skill acquisition: Learning and performance effects reconsidered. *Research Quarterly for Exercise and Sport*, *59*(4), 277–287. [CrossRef]

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, *25*(3), 639–647. [CrossRef]

Lonergan, R., Cumming, T. M., & O'Neill, S. C. (2022). Exploring the efficacy of problem-based learning in diverse secondary school classrooms: Characteristics and goals of problem-based learning. *International Journal of Educational Research*, *112*, 101945. [CrossRef]

* Lotfolahi, A. R., & Salehi, H. (2017). Spacing effects in vocabulary learning: Young EFL learners in focus. *Cogent Education*, *4*(1), 1287391. [CrossRef]

Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. S. (2019). How the amount and spacing of retrieval practice affect the short-and long-term retention of mathematics knowledge. *Educational Psychology Review*, *32*(1), 277–295. [CrossRef]

* Lyle, K. B., Bego, C. R., Ralston, P. A., & Immekus, J. C. (2022). Spaced retrieval practice imposes desirable difficulty in calculus learning. *Educational Psychology Review*, *34*(3), 1799–1812. [CrossRef]

Maddox, G. B. (2016). Understanding the underlying mechanism of the spacing effect in verbal learning: A case for encoding variability and study-phase retrieval. *Journal of Cognitive Psychology*, *28*(6), 684–706. [CrossRef]

* McNellis, C. J. (2015). Re-conceptualizing instruction on the statement of cash flows: The impact of different teaching methods on intermediate accounting students' learning. *Advances in Accounting Education*, *17*, 115–144.

Microsoft Corporation. (2023). *Microsoft excel*. (Version 2032). Microsoft Corporation.

* Moss, V. D. (1996). The efficacy of massed versus distributed practice as a function of desired learning outcomes and grade level of the student. *Dissertation Abstracts International Section B: The Sciences and Engineering*, *56*(9-B), 5204. Available online: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc3&NEWS=N&AN=1996-95005-375 (accessed on 16 May 2023).

Namaziandost, E., Murad Hassan Mohammed, S., & Masoumeh Izadpanah, S. (2020). The effects of spaced versus massed distribution instruction on EFL learners' vocabulary recall and retention. *Cogent Education*, *7*(1), 1792261. [CrossRef]

Namaziandost, E., Nasri, M., Fariba Rahimi, E., & Mohammad Hossein, K. (2019). The impacts of spaced and massed distribution instruction on EFL learners' vocabulary learning. *Cogent Education*, *6*(1), 1661131. [CrossRef]

Namaziandost, E., Razmi, M. H., Atabekova, A., Shoustikova, T., & Kussanova, B. H. (2023). An account of Iranian intermediate EFL learners' vocabulary retention and recall through spaced and massed distribution instructions. *Journal of Education*, *203*(2), 275–284. [CrossRef]

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, 71. [CrossRef] [PubMed]

Pappas, M. L. (2009). Inquiry and 21st-century learning. *School Library Monthly*, *25*(9), 49–51.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contourenhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, *61*(10), 991–996. [CrossRef]

* Petersen-Brown, S., Lundberg, A. R., Ray, J. E., Dela Paz, I. N., Riss, C. L., & Panahon, C. J. (2019). Applying spaced practice in the schools to teach math vocabulary. *Psychology in the Schools*, *56*(6), 977–991. [CrossRef]

* Petersen-Brown, S., Riese, A. M., Schneider, M. M., Ray, J. E., & Clonkey, H. R. (2023). Applying distributed practice in the schools to enhance retention of spelling words. *Psychology in the Schools*, *60*(9), 3486–3503. [CrossRef]

Pomerance, L., Greenberg, J., & Walsh, K. (2016). *Learning about learning: What every new teacher needs to know*. National Centre on Teacher Quality. Available online: https://www.nctq.org/publications/Learning-About-Learning:-What-Every-New-Teacher-Needs-to-Know (accessed on 12 April 2023).

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 20 April 2023).

* Reynolds, J. H., & Glaser, R. (1964). Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educational Psychology*, *55*(5), 297–308. [CrossRef]

Robinson, A. G. (1979). A compared difference in pupil performance as a function of teacher presented material (1-1), peer (cross-age) presented material, teacher presented material (1-5), and massed and spaced scheduling of instruction. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *39*(11-A), 6701–6702. Available online: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc2&NEWS=N&AN=1980-51592-001 (accessed on 16 May 2023).

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher*, *39*(5), 406–412. [CrossRef]

Saavedra, A. R., & Opfer, V. D. (2012). Learning 21st-century skills requires 21st-century teaching. *The Phi Delta Kappan*, *94*(2), 8–13. [CrossRef]

* Sayeski, K. L., Earle, G. A., Eslinger, R., & Whitenton, J. N. (2017). Teacher candidates' mastery of phoneme-grapheme correspondence: Massed versus distributed practice in teacher education. *Bulletin of the Orton Society*, *67*(1), 26–41. [CrossRef]

Schopfel, J., & Rasuli, B. (2018). Are electronic theses and dissertations (still) grey literature in the digital age? A FAIR debate. *The Electronic Library*, *36*(2), 208–219. [CrossRef]

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*(5), 763–767. [CrossRef]

Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rucker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*, d4002. [CrossRef] [PubMed]

Suzuki, Y., & Dekeyser, R. (2017). Exploratory research on second language practice distribution: An Aptitude × Treatment interaction. *Applied Psycholinguistics*, *38*(1), 27–56. [CrossRef]

Svihla, V., Wester, M. J., & Linn, M. C. (2017). Distributed practice in classroom inquiry science learning. *Learning: Research and Practice*, *4*(2), 180–202. [CrossRef]

Sweller, J. (2010). Cognitive load theory: Recent theoretical advances. In J. L. Plass, R. Brünken, & R. Moreno (Eds.), *Cognitive load theory* (pp. 29–47). Cambridge University Press.

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*(6), 837–848. [CrossRef]

The EndNote Team. (2023). *EndNote*. (Version 20.6). Clarivate.

van Merrienboer, J. J. G., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational Psychologist*, *38*(1), 5–13. [CrossRef]

Viechtbauer, W. (2010). Conducting meta-analyses in {R} with the {metafor} package. *Journal of Statistical Software*, *36*(3), 1–48. [CrossRef]

* Weaver, J. R. (1976). The relative effects of massed versus distributed practice upon the learning and retention of eighth grade mathematics. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *37*(5-A), 2698.

Wilson, D. B. (n.d.). *Practical meta-analysis effect size calculator*. Available online: https://campbellcollaboration.org/research-resources/effect-size-calculator.html (accessed on 16 June 2023).

Wiseheart, M., Küpper-Tetzel, C. E., Weston, T., Kim, A. S. N., Kapler, I. V., & Foot-Seymour, V. (2019). Enhancing the quality of student learning using distributed practice. In J. Dunlosky, & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 550–584). Cambridge University Press.

* Yazdani, M., & Zebrowski, E. (2006). Spaced reinforcement: An effective approach to enhance the achievement in plane geometry. *Journal of Mathematical Sciences and Mathematics Education*, *1*, 37–43.

Ysseldyke, J. E., & Pickholtz, H. (1975). Dissertation research in school psychology: 1967–73. *Journal of School Psychology*, *13*(3), 264–271. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Students' Perceptions of Effective Math Learning Strategies

Marissa K. Hartwig [1,*] and Doug Rohrer [2]

[1] Department of Psychology, University of Maryland, College Park, MD 20742, USA
[2] Department of Psychology, University of South Florida, Tampa, FL 33620, USA; drohrer@usf.edu
* Correspondence: mhartwig@umd.edu

Abstract

Two highly effective math learning strategies are *spaced practice* (in which problems of the same kind are distributed across many sessions) and *interleaved practice* (in which problems of different kinds are mixed rather than blocked). Though these strategies are supported by data, students may be reluctant to use them if they perceive the strategies as ineffective or unpleasant. In Study 1, we surveyed 174 grade 7 math students about the efficacy and likability of spaced and interleaved practice. Spaced practice was often judged likable, but nearly half of students failed to recognize its efficacy. Interleaved practice was judged both unlikable and inefficacious by most students. In Study 2, we further explored perceptions of interleaving in a survey of 233 grade 7 math students. Again, students erroneously judged interleaved practice to have low efficacy. Compared to blocked practice, interleaved practice was judged less effective, less preferable, more time-consuming, and more difficult. This work identifies perceptions that may discourage students from using effective learning strategies and also shows that specific perceptions differ by strategy. Helping students overcome their negative perceptions of spacing and interleaving is an important future direction.

Keywords: math learning; spaced practice; interleaved practice; student perceptions; metacognition

1. Introduction

Considerable research has identified learning strategies that are reliably effective across various learning conditions, tasks, and materials (for a review, see Dunlosky et al., 2013), but much less is known about students' perceptions of these strategies, which can affect whether students use them. Two highly effective strategies for math learning are *spaced practice* (in which problems of the same kind are distributed across multiple sessions) and *interleaved practice* (in which problems of different kinds are mixed rather than blocked). The evidence for these two strategies is strong, as reviewed below, but empirical efficacy and expert recommendation is not sufficient to ensure utilization. Students may be reluctant to use strategies they perceive as ineffective, difficult, or unpleasant, and even teachers may be disinclined to impose strategies they know their students dislike. Thus, in two studies presented here, we examined students' perceptions of spaced and interleaved math practice.

1.1. Evidence for Spaced and Interleaved Practice

Both spaced practice and interleaved practice are supported by a wealth of empirical evidence and have been widely recommended for use in classrooms (e.g., Carpenter, 2014; Carpenter et al., 2012; Kang, 2016; Roediger & Pyc, 2012; Rohrer, 2012, 2015; Weinstein

et al., 2018). Here we provide only a brief overview of the evidence, as numerous authors have published reviews pertaining to the efficacy of spaced and interleaved[1] practice (e.g., E. L. Bjork & Bjork, 2011; Carpenter et al., 2012, 2022; Cepeda et al., 2006; Dunlosky et al., 2013; Firth et al., 2021; Kang, 2016; Roediger & Pyc, 2012; Son & Simon, 2012). Though these strategies can support learning in many different content areas, most important for the present research is that these strategies have been demonstrated to be highly effective for math learning, as we describe next.

In math spacing experiments, practice problems of one kind (e.g., adding fractions) are either distributed more widely (e.g., across 2 weeks) or concentrated into a shorter period (e.g., across 2 days), and results typically show that final test scores are better when practice is more distributed (spaced), despite equating the number of practice problems and test delay. The efficacy of spaced practice for math problem solving has been demonstrated in laboratory experiments (e.g., Gay, 1973; Rohrer & Taylor, 2006, 2007) and classroom studies (e.g., Barzagar Nazari & Ebersbach, 2019; Emeny et al., 2021) with students ranging from grade schoolers (Schutte et al., 2015) to college students (Hopkins et al., 2016; Lyle et al., 2020) and with a variety of math topics (e.g., geometry, algebra, diagrams, permutations, pre-calculus; see Carpenter et al., 2022). According to various theories of the spacing effect (for a review, see Delaney et al., 2010), spaced practice gives students a mental break that permits renewed attention in subsequent sessions and memory consolidation in the interim, as well as promoting varied memory cues in different learning contexts and retrieval of information from earlier sessions—all of which can support learning and retention of studied materials. No single explanatory theory of spacing is agreed upon, and multiple mechanisms may operate concurrently, but the benefit of spacing is undisputed. To achieve durable learning, math assignments should be spaced such that learners must revisit and reapply math concepts and strategies across time.

When math practice is interleaved, students practice a mix of different problem types within the same assignment (e.g., adding fractions, solving proportions, and computing slope), compared to assignments consisting of blocks of one type (e.g., all slope problems). Interleaved math practice has been shown to reliably produce better test scores than blocked practice in laboratory experiments (e.g., Mayfield & Chase, 2002; Rohrer & Taylor, 2007) and classroom studies (e.g., Rohrer et al., 2015, 2020b) with students of different ages (e.g., Foster et al., 2019; Taylor & Rohrer, 2010) with topics of varying complexity and similarity (e.g., Rohrer et al., 2014; Sana et al., 2017) and at varying test delays (e.g., Rau et al., 2013; Ziegler & Stern, 2014). Interleaved math practice helps students learn to identify appropriate strategies for each problem, rather than merely repeating the procedure of the preceding problem; it also inherently increases the spacing of each problem type and encourages students to try to retrieve the needed information (e.g., fact, formula, procedure) from memory (see Rohrer et al., 2020b, for further discussion). In short, interleaved math practice entails a combination of strategy selection, spacing, and retrieval that yields sizable benefits for students' test performance.

Spaced and interleaved math practice are strongly supported by evidence and recommended for use in math classrooms and textbooks (Carpenter, 2014; Dunlosky, 2013; Pan, 2015; Roediger & Pyc, 2012; Rohrer et al., 2020a), but recommendations alone may not be enough. Whether strategies get used in classrooms can, in part, be determined by the experiences that students and teachers have with those strategies. Do the strategies feel effective? How likable are they? If a strategy seems ineffective or unpleasant for students, then students (or their teachers) might choose other—possibly less effective—strategies that students find preferable.

1.2. Perceptions of Spaced and Interleaved Practice

Student perceptions of learning strategies influence which strategies they use. Models of self-regulated learning predict that student metacognitive knowledge or beliefs about learning activities affect study strategy decisions and motivation (e.g., Zimmerman, 2000; for a review of models, see Panadero, 2017). Unfortunately, explicit instruction about effective learning strategies is rare in classrooms, instead leaving students to rely on prior experience, intuitions, and heuristics—all of which can produce flawed inferences. Students' experiences with learning strategies do not typically occur under isolated and controlled conditions that would permit accurate comparisons of efficacy. Students' attempts to judge strategy efficacy may be further thwarted when learning outcomes are not immediate or if the student has previously succeeded despite using poor strategies. How a strategy feels while using it can also lead students astray: when effective study strategies are perceived to require more effort, students may perceive them to be less efficacious or efficient (R. A. Bjork et al., 2013; Rea et al., 2022). Students may rely on other shortcuts as well to judge strategy efficacy (e.g., how their peers study), which can yield poor guidance. In a framework for improving student self-regulation of learning strategies, McDaniel and Einstein (2020) proposed conditions needed to support students' use of effective strategies, including the belief that a strategy is effective, knowledge about how to use it, commitment to using it, and a plan to implement it. Undoubtedly, students' perceptions of strategy efficacy are fundamental to fostering the use of effective strategies.

Do students perceive spacing and interleaving to be efficacious for learning? Surveys of college students' study behaviors and beliefs provide insight. With respect to spacing, 85% of college students said spaced study was better than massed study for long-term retention of material (Susser & McCabe, 2013), and 81% said flashcards should be spaced rather than massed (Wissman et al., 2012). When asked to judge a vignette comparing spaced and massed practice, 69% of college students judged spaced practice to be more effective (Morehead et al., 2016). This evidence suggests that students have some awareness of the efficacy of spaced practice. However, though many students may correctly identify spaced practice as more effective than massed practice, they may not recognize the benefit of practice that is more spaced versus less spaced. For instance, when asked to depict ideal spacing or to hypothetically schedule study activities to maximize exam performance (e.g., Blasiman et al., 2017; Cohen et al., 2013; Hartwig et al., 2022; Taraban et al., 1999; Wissman et al., 2012), many college students opted for study to be concentrated (i.e., less spaced) near exams rather than spacing those same study activities across a longer time period.

With respect to interleaving, perceptions of efficacy may be especially lacking. In lab studies, participants have been asked to judge interleaving efficacy after using both interleaved and blocked practice to learn to classify novel paintings by artist style (e.g., Kornell & Bjork, 2008; Kornell et al., 2010; Yan et al., 2016). In these studies, many participants incorrectly believed that blocked practice had been more effective than interleaved—even though their own test performance showed the opposite—and were not easily convinced of interleaving's superiority (Yan et al., 2016). In other studies, when given a vignette describing interleaved and blocked practice for the artist learning task, only 16% of college students correctly judged interleaved practice to be more effective (Morehead et al., 2016; see also Halamish, 2018; McCabe, 2011). The perceived superiority of blocked practice is not limited to the artist learning paradigm. Studies involving hypothetical scheduling of exam preparation have also shown that participants commonly select blocked over interleaved practice (Hartwig et al., 2022; Yan et al., 2017; Yan & Sana, 2021).

Even in a familiar subject area like mathematics, in which K-12 and college curricula commonly involve revisiting concepts and skills across time (spacing) and reviewing a mix of problem types before exams (interleaving), research suggests that students still

undervalue these strategies. In one recent study (Hartwig et al., 2022), college students were asked to create or select hypothetical 2-week schedules that they believed would maximize performance on a final math test. The authors then computed the degree of spaced and interleaved practice in those schedules and found substantially less spacing and interleaving than would be optimal for learning. To the degree that participants' schedules provided spacing or interleaving, it was mostly due to the inclusion of a pre-exam review session intended to refresh one's memory on each problem type shortly before the test, rather than a sophisticated understanding of the efficacy of spacing practice across time or interleaving problem types within sessions. In short, spaced and interleaved practice are mistakenly judged to have low efficacy in many contexts, including math learning.

Student enjoyment of learning strategies may also be an important perception that can affect their strategy choices, though this has not been the subject of much previous investigation. There is some evidence that students choose learning strategies they find more enjoyable (e.g., Obergriesser & Stoeger, 2020). Further, it is plausible that the expected enjoyment (or perceived likability) of a learning strategy may be related to other perceptions—such as the perceived efficacy or perceived difficulty of the strategy. How (or whether) such perceptions are related to one another and shape students' preferences for spaced and interleaved math practice remains largely unexplored—but could provide insights about students' willingness to use these highly effective strategies.

Of course, the role of students in deciding which strategies they use can differ by situation and change over time, with students typically exercising more control as they advance in school and eventually, as young adults, making many study decisions on their own. In the present studies, we sampled students in middle school—an age at which students may develop learning habits and strategy perceptions that persist into high school and college. Knowing middle schoolers' perceptions might allow early targeting of interventions, possibly improving academic outcomes over a longer trajectory. Certainly, teachers play an important role in shaping students' strategy use, if not through direct instruction about learning strategies, then via the lessons, assignments, and other work or activities they plan. Though such plans are teacher-generated, teachers are not impervious to students' opinions and may plan tasks they believe their students will perceive positively. Students' perceptions of these tasks can affect their engagement in class and also which strategies they adopt when studying on their own. Whether by teachers' decisions or students' own decisions, student perceptions of learning strategies can influence which strategies get used.

1.3. Overview of the Present Studies

In the present two studies, we surveyed middle school students in their math classes about their perceptions of two highly effective math learning strategies—spaced practice and interleaved practice—in the context of hypothetical math assignments. In Study 1, students judged the efficacy and likability of the two strategies. For comparison, we also surveyed their perceptions of more (vs. less) practice, because we suspected that most students would recognize *more practice* as beneficial, if sometimes unlikable. To foreshadow, Study 1 raises concerns about students' perceptions of interleaved practice, so Study 2 further addressed perceptions of interleaving. Along with perceptions of efficacy and preference, Study 2 examined other perceptions that could help explain why students dislike and fail to recognize the efficacy of interleaved practice, such as perceptions of difficulty, interest, and the time required for interleaved (vs. blocked) practice.

2. Study 1

While the efficacy of spacing and interleaving may not always be recognized by students, the strategy of doing practice problems (i.e., better to do more practice than less

practice) is intuitive and likely to be viewed as efficacious. Thus, the perceived efficacy of more (vs. less) practice can provide a benchmark for comparison with students' perceptions of spaced practice and interleaved practice. We suspected that most students would judge more (vs. less) practice to be efficacious, whereas fewer students would recognize the efficacy of spacing and interleaving. We also asked students to judge strategy likability, which is previously unexplored for these math learning strategies. We suspected that the strategies might differ in their likability, though we did not know which strategies would be most or least likable. Importantly, likability adds to the picture of students' willingness to use each strategy.

2.1. Method

2.1.1. Participants

The participants were 174 students in 7th grade math (ages 12–13) at a middle school in a large school district in Florida. The math course was the modal course for grade 7 students at most schools in the district. We sampled eight class sections of the course taught by two teachers (four sections each). The sample size was large enough to give margins of error smaller than 7.5% on estimated proportions of students, at 95% confidence.

2.1.2. Materials

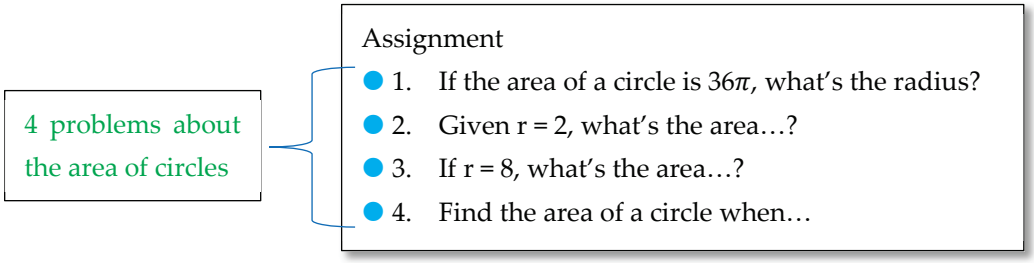
Math practice worksheet. To give students recent experience with interleaved and blocked practice, we created a practice worksheet consisting of 8 math problems (4 blocked and 4 interleaved). Practice problems required students to compute the volume of a cylinder, use the Pythagorean theorem, solve an equation for x, compute the volume of a sphere, and identify equivalent mathematical expressions. We also created an answer key showing worked solutions, which we provided to the teachers. Students completed the practice worksheet before taking the survey.

Survey booklet. A survey booklet presented three math learning scenarios. The first scenario addressed *more practice*—a strategy we expected students would recognize as efficacious. The second and third scenarios addressed the two strategies of primary interest—*spaced practice* and *interleaved practice*, respectively. Each scenario was presented on its own page and included a short explanation of the scenario followed by two options (labeled "A" and "B")—a more efficacious option (e.g., more spacing) and a less efficacious option (e.g., less spacing), with the order of the two options counterbalanced (i.e., more efficacious vs. less efficacious option shown first). The options were presented as diagrams like in Figures 1–3 (but with "Option A" or "Option B" in place of the underlined labels). In the first scenario, students were told to imagine learning a new skill (e.g., finding the area of a circle) followed by two options (Figure 1): a 4-problem assignment or an 8-problem assignment. In the second scenario, students considered two options for the spacing of slope problems (Figure 2): one day apart or three days apart. In the third scenario, students considered two options for interleaving (Figure 3): 6 problems of the type learned that day or 3 problems learned that day plus 3 problems of previously-learned types (where both options included an interleaved review on Day 7 to equate the recency of each problem type and to increase realism).

For each scenario, students judged (1) perceived efficacy ("Which option would be *more helpful* to prepare you for a test… given a month later?) and (2) likability ("Which option would you *like more*?) by selecting among 5 responses that corresponded to the chosen option (A or B) and strength of the choice (much more, slightly more, equal). The order of the two judgments (i.e., efficacy, likability) was counterbalanced. The complete survey with instructions and administration details, as well as the data for both studies, can be found at https://osf.io/rwbq4/ (accessed on 3 June 2025).

Less Practice

The teacher assigns **4 problems** to practice the new skill.



More Practice (recommended strategy)

The teacher assigns **8 problems** to practice the new skill.

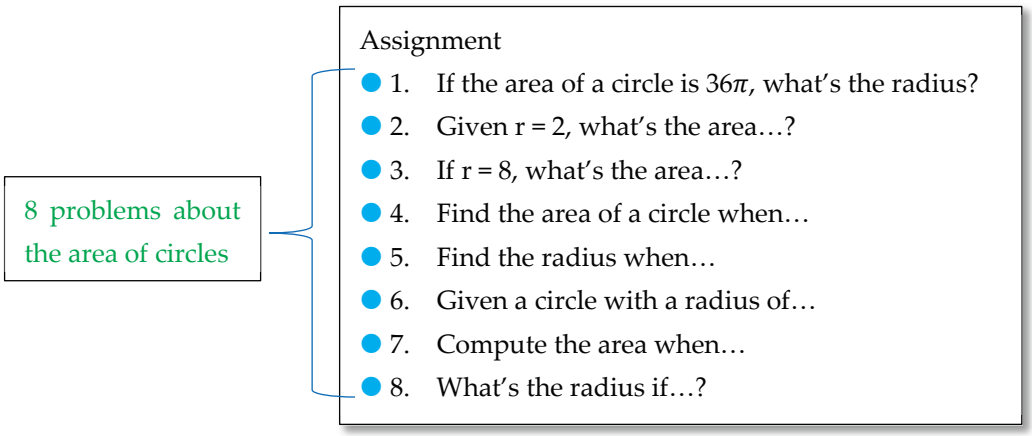
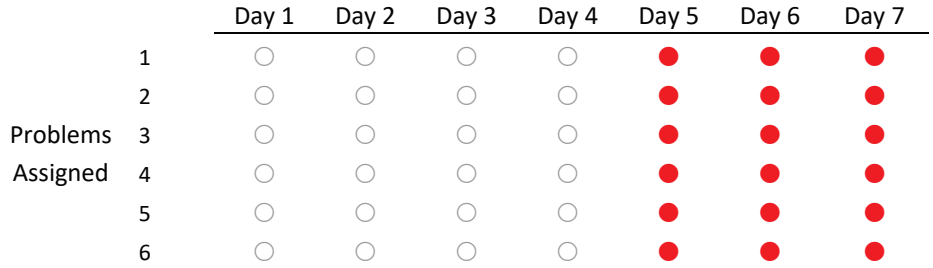


Figure 1. Less Versus More Practice (Study 1). The green words summarize each hypothetical assignment. The blue dots signify that each problem addresses the same new skill (area of circles).

Less Spacing

Slope problems (red dots) are completed **1 day apart**.



More Spacing (recommended strategy)

Slope problems (red dots) are completed **3 days apart**.

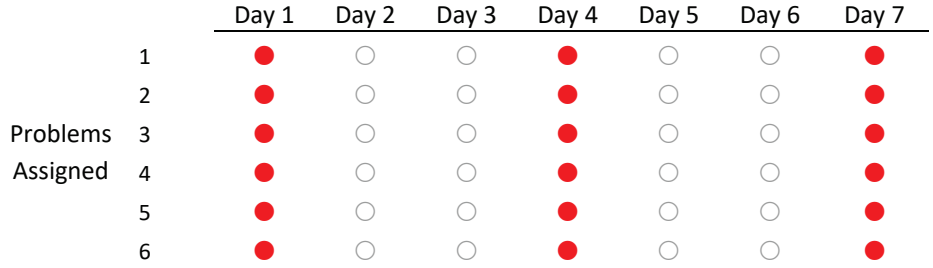


Figure 2. Less Versus More Spacing (Study 1). Red dots represent slope problems and white dots represent various other types of math problems.

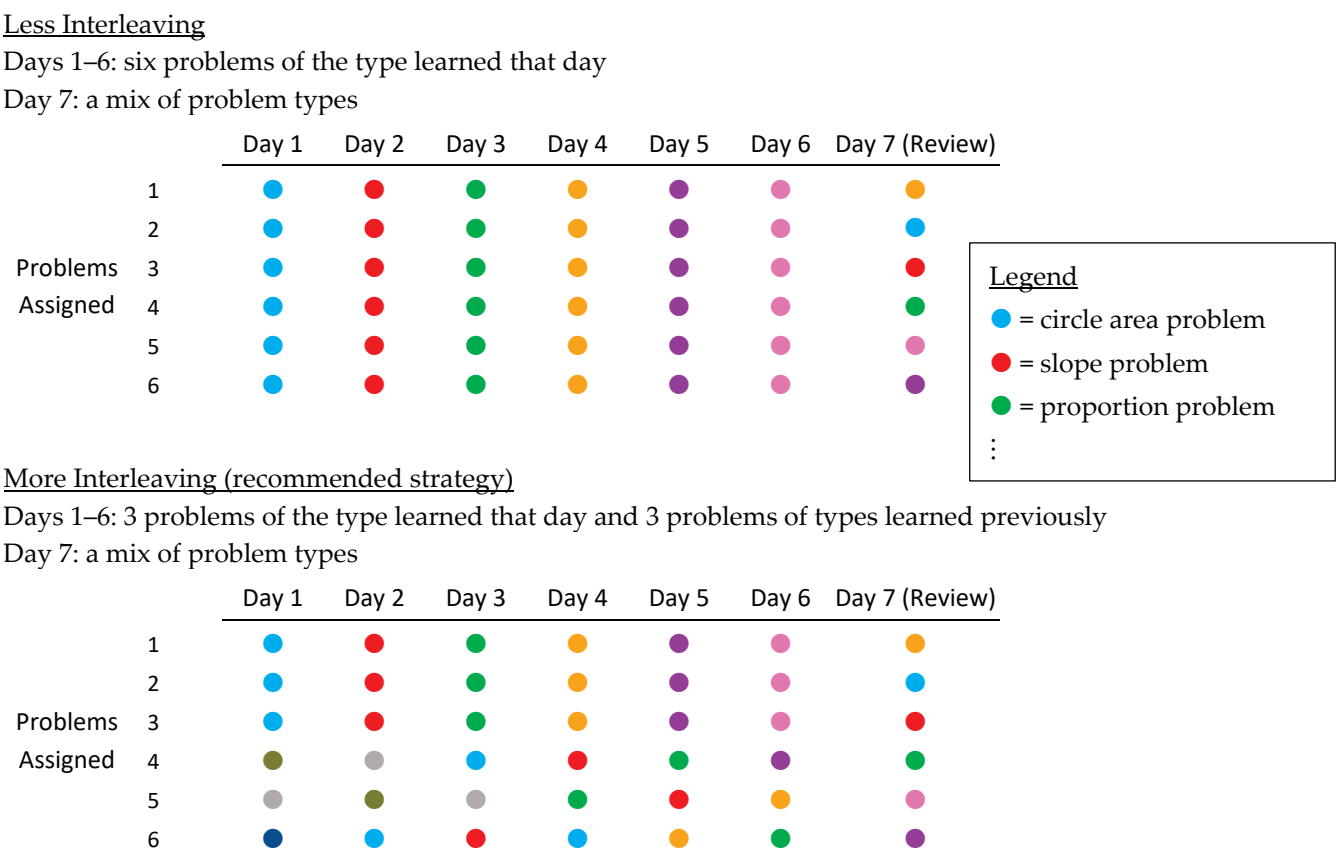


Figure 3. Less Versus More Interleaving (Study 1). Each dot color represents a different type of math problem (see Legend for examples).

2.1.3. Procedure

The study occurred during a class period without the researchers present. The teacher distributed the math practice worksheet to each student in the participating class. Students were instructed to try solving the practice problems and were permitted to work together and seek help if needed. When most students were finished or no longer making progress on the problems, the teacher then distributed a survey booklet to each student and provided an oral explanation (prepared by the researchers) that guided students through each page. The oral explanation for each scenario explained to students what the various dots represented (e.g., in the spacing scenario, red dots represented slope problems and white dots represented various other types of math problems; in the interleaving scenario, example topics were given for different dot colors). After students completed the survey, the teacher collected the booklets. Finally, the teacher presented the solutions to the math problems on the practice worksheet. The entire procedure (i.e., practice worksheet, survey, and problem solutions) required approximately 40 min and was completed within one class period. The survey booklets were returned to the researchers, but the practice worksheets were not.

2.2. Results and Discussion

Students' perceptions of the efficacy of each strategy are shown in the left column of Figure 4. For the strategy of doing more practice (vs. less practice), 69% of students said the option with more problems was slightly or much more effective than the alternative. In other words, most (but not all) students recognized the efficacy of doing practice problems—approximately as expected. For the strategy of spacing, however, students' perceptions of efficacy were more varied, with responses relatively evenly distributed

across the five response choices, indicating that students' insight into the efficacy of spacing was limited at best. Chi Square goodness-of-fit tests were used to determine whether the response pattern for each survey question differed from random responding, and only the question about spacing efficacy was indistinguishable from random (see OSF for details), which could reflect students' differing opinions or uncertainty about spacing efficacy. Finally, for the strategy of interleaving, only 22% said the option with more interleaving was slightly or much more effective, whereas 65% said the option with little interleaving (i.e., mostly blocked) was more effective, contrary to actual efficacy. Thus, most students had erroneous perceptions about the efficacy of interleaving.

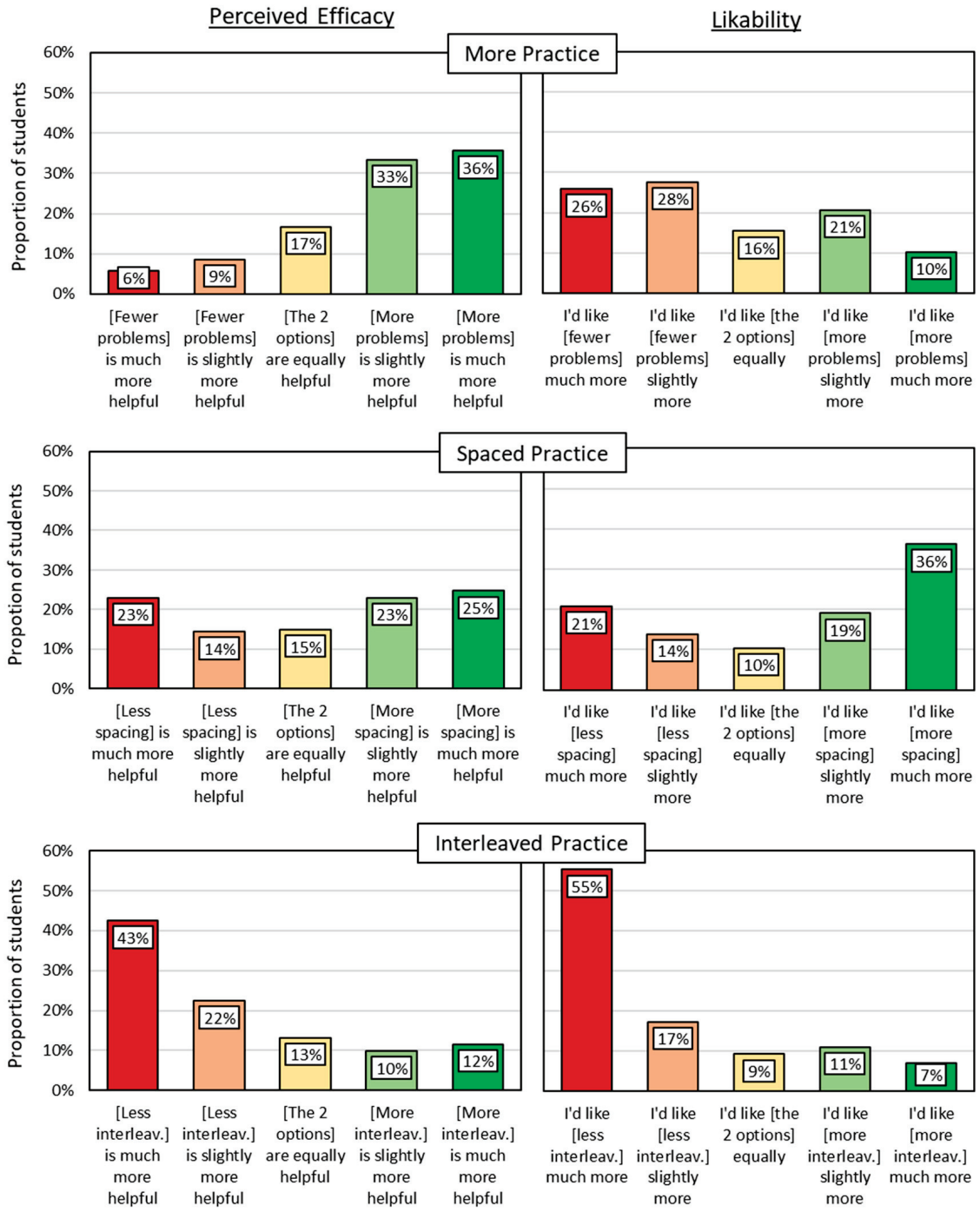


Figure 4. Perceived Efficacy and Likability of Math Learning Strategies (Study 1).

Students' perceptions of likability of each strategy are shown in the right column of Figure 4. Regarding the quantity of practice problems, more than half of students preferred the less efficacious strategy of doing fewer (rather than more) problems. Regarding spacing, however, more than half of students preferred the more efficacious strategy of more spacing (rather than less spacing). Finally, regarding interleaving, students overwhelmingly (72%) preferred the less efficacious strategy of mostly blocked practice (rather than interleaved practice).

In summary, the middle school students in Study 1 were mostly aware that doing practice problems helps their learning, even if they do not always like it. Concerning spaced practice, many students found it likable, yet they had no consensus about its efficacy. This result signals an opportunity to educate students about the efficacy of spacing which, given its likability, may face little resistance from students. The results for interleaving, however, were more troubling. Students often erroneously believed that learning was benefitted by less (rather than more) interleaving, and this misperception about interleaving efficacy was paired with low likability. These results paint a grim picture regarding students' inclination for interleaved practice. In Study 2, we surveyed additional perceptions that could help explain the low perceived efficacy and low likability of interleaved practice. For example, students might perceive interleaved practice to be more difficult or time-consuming than blocked practice, and such perceptions might be interpreted by students as evidence of low efficacy (E. L. Bjork & Bjork, 2011; Kirk-Johnson et al., 2019) and contribute to low likability. As in Study 1, we again relied on diagrams to help students understand the strategy in question, avoiding words like "interleaving" that may be confusing or unfamiliar to them. Unlike Study 1, however, Study 2 used an alternative diagram offering a simpler depiction of interleaved and blocked practice worksheets.

3. Study 2

Students' negative perceptions of interleaved practice—as unlikable and inefficacious—are concerning and merit additional investigation, so in Study 2 we further explored perceptions of interleaved practice. We again surveyed middle school math students, but we used a different illustration to depict interleaved ("mixed") practice and its alternative, blocked practice (see Figure 5). Using this illustration, we queried students about the efficacy of the two alternatives, students' preference[2] between them, the time required to complete them, how difficult they are, and how interesting they are.

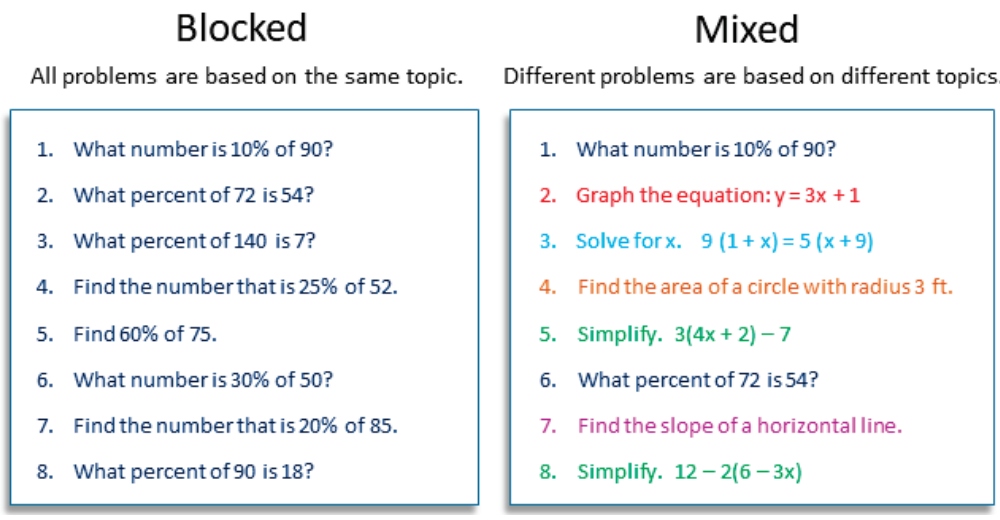


Figure 5. Example of Blocked versus Interleaved (Mixed) Practice (Study 2).

3.1. Method

3.1.1. Participants

Participants were 233 students in 7th grade math (ages 12–13) at a middle school in a large school district in Florida. The school was the same as in Study 1 but during a different school year, and the two studies did not include any of the same participants. The math course was the modal course for grade 7 students at most schools in the district. We sampled twelve class sections of the course taught by four teachers (two to four sections each). The sample size was large enough to give margins of error smaller than 6.5% on estimated proportions of students, at 95% confidence.

The survey reported here was part of a larger, classroom-based, full-year study that ended unexpectedly, before we could administer the classroom-based final test, when the coronavirus pandemic forced schools to close abruptly and move all instruction online for the remainder of the school year. In its place, the participating teachers agreed to an online survey reported here. The original study involved an in-person, within-subjects manipulation of math practice in the classroom. It was not expected to affect survey responses and occurred 8 weeks before the survey was administered.

Of the 260 students in the original study, 234 completed the online survey (90%). One participant was excluded from analysis for spending less than 10 s on the entire survey, which we believe was too little time to examine the diagram, read the survey questions, and give meaningful responses. Our final sample consisted of 233 students.

3.1.2. Materials and Procedure

A survey link was given to teachers who distributed it to the students in their participating classes. Students could complete the survey at any time during the week it was assigned. The survey took less than 5 min and was followed by an online math task unrelated to the survey. The survey presented students with a diagram of hypothetical interleaved and blocked assignments (see Figure 5), along with a brief description of each. With the diagram still visible, students were asked to make five comparisons of the two assignments: time ("Which kind of assignment takes you more time to finish?"), difficulty ("...is harder for you?"), interest ("...is more interesting?"), efficacy ("...helps you score higher on [standardized test[3]]?"), and preference ("...do you prefer?"). (Response options for each question are visible in Figure 6, and a copy of the survey is available at the OSF link provided above.) As a reward for participation, participants received a small gift bag of school supplies at the end of the school year.

3.2. Results and Discussion

Students' survey responses are shown in Figure 6. Like in Study 1, most students in Study 2 failed to recognize the greater efficacy of interleaved practice over blocked practice: only 30% of students said that interleaved practice is slightly or much more effective than blocked, whereas 52% said the opposite. Students also reported a clear preference for blocked practice (65% preferring blocked vs. 18% preferring interleaved). Regarding interest, assignment type did not have a consistent impact: 34% said blocked assignments are more interesting, 33% said interleaved assignments are more interesting, and 34% said they are about the same. Most students agreed, however, that interleaved assignments require more time (74%) and are more difficult (63%).

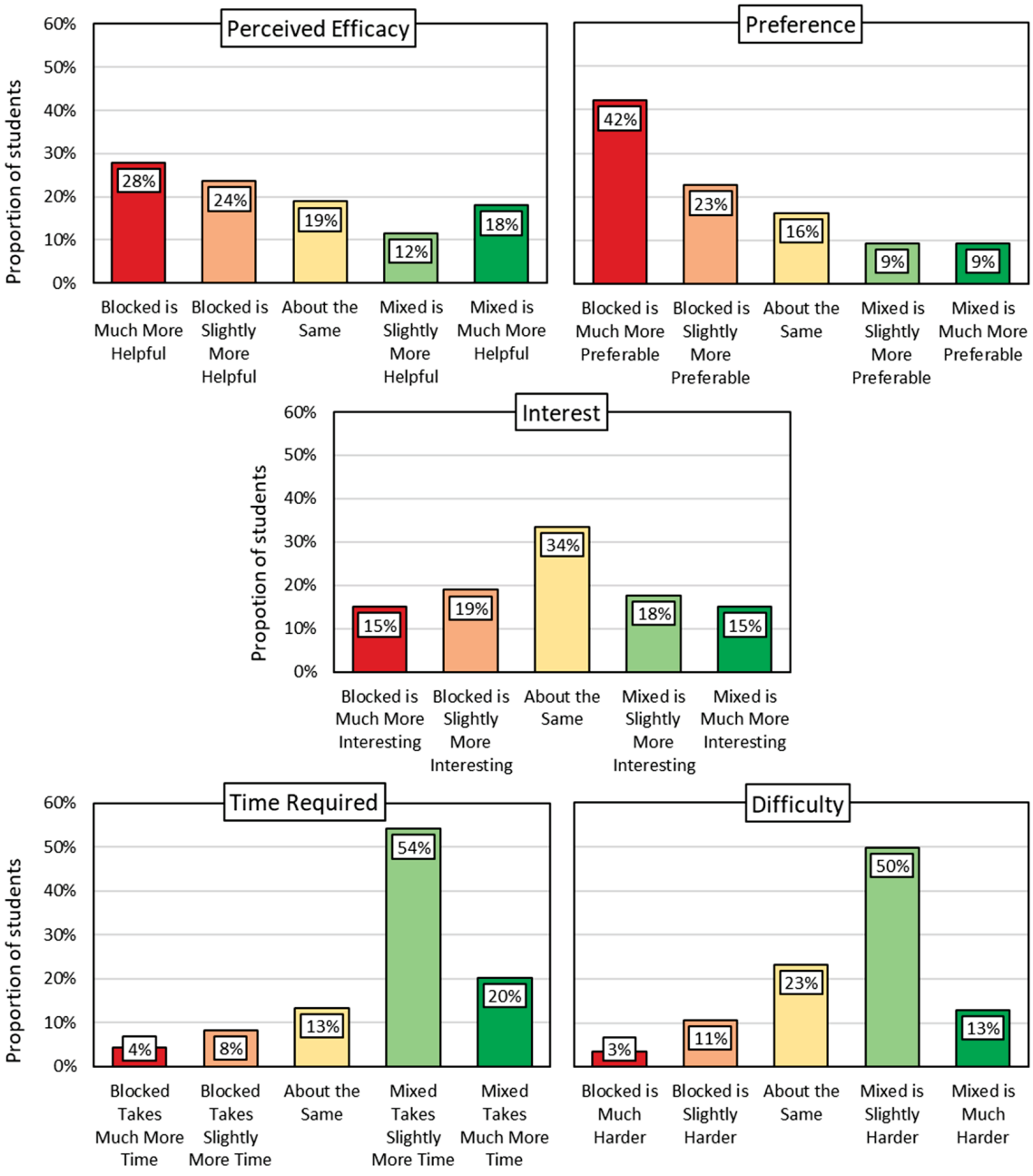


Figure 6. Blocked vs. Interleaved Practice (Study 2).

We also examined the associations among the five perceptions surveyed (Table 1). Students who believed that interleaved assignments take more time (than blocked practice) also generally believed that interleaved assignments are more difficult ($r = 0.66$). Interestingly, perceiving interleaved practice as more time-consuming (less efficient) and more difficult was more strongly associated with low preference than with low perceived efficacy. In other words, although difficulty may indeed be interpreted by students as signaling low strategy efficacy, difficulty may diminish preference even more. Students who more strongly preferred interleaved practice (over blocked) viewed it as less difficult ($r = -0.61$) and less time consuming ($r = -0.57$) but more interesting ($r = 0.51$) and more effective ($r = 0.48$). Altogether, these results suggest that to encourage the use of interleaved practice, researchers and educators must not only convince students of its efficacy but also find ways to overcome negativity pertaining to difficulty, inefficiency, and unlikability.

Table 1. Correlations Among Perceptions of Interleaved Practice (Study 2).

Perceptions	1	2	3	4	5
1. Time Required	–				
2. Difficulty	0.66 *	–			
3. Interest	−0.30 *	−0.41 *	–		
4. Efficacy	−0.30 *	−0.28 *	0.33 *	–	
5. Preference	−0.57 *	−0.61 *	0.51 *	0.48 *	–

* $p < 0.01$.

4. General Discussion

In the studies presented here, we investigated middle school students' perceptions of spaced and interleaved practice—two highly effective math learning strategies. In Study 1, spaced practice was often judged likable, but nearly half of students failed to recognize its efficacy, whereas interleaved practice was judged both unlikable and inefficacious by most students. In Study 2, further exploration of students' perceptions of interleaved practice pointed to possible reasons that most students dislike interleaving and fail to perceive its efficacy—including perceptions that interleaved assignments are time-consuming and difficult. Understanding students' perceptions of learning strategies is important because these perceptions can influence whether the strategies get used (Zepeda et al., 2020). The present findings highlight the need to consider not only students' metacognitive knowledge of strategy efficacy but also strategy likability.

With respect to students' perceptions of strategy efficacy, the present work suggests that inadequate metacognitive knowledge may be a barrier for both spaced and interleaved practice, albeit to different degrees. When students compared more spaced versus less spaced practice, they lacked a consensus regarding efficacy, suggesting that, for at least half the students, their strategy knowledge needs improvement. The strategy knowledge of these 7th graders might, at first glance, seem deficient compared to college students who recognize the efficacy of *spaced over massed* practice (e.g., Susser & McCabe, 2013), but recognizing the efficacy of *more spacing over less spacing* is difficult for college students also (e.g., Hartwig et al., 2022). With interleaved practice, perceptions of efficacy were even more concerning: most students (65%) erroneously believed that less interleaved (mostly blocked) practice was superior to more interleaved practice (Study 1), and even a clear-cut comparison of fully blocked vs. fully interleaved assignments (Study 2) found that more than half of the students perceived blocked to be superior. This is consistent with prior work showing that students commonly draw the wrong conclusion about interleaved vs. blocked efficacy (e.g., Kornell & Bjork, 2008) and that dissuading them from this misperception is difficult (Yan et al., 2016). How to best remediate students' erroneous metacognitive beliefs about spacing and interleaving should be studied in future work to provide a basis for successful intervention (see McDaniel & Einstein, 2020, for further discussion).

The likability of a strategy also affects its use (Obergriesser & Stoeger, 2020), and in the present study, likability was shown to differ considerably by strategy. Spacing was judged favorably, with more than half of students in Study 1 reporting greater liking for practice that is more spaced rather than less spaced. This suggests that likability is not a large barrier for spaced practice. In contrast, the likability of interleaved practice does appear to be a substantial hurdle, with 72% of students liking practice that is less interleaved (vs. more interleaved; Study 1) and 65% preferring blocked assignments (vs. interleaved assignments; Study 2). Finding ways to increase the likability of interleaved practice is an important future direction to encourage its use.

We also explored other potential barriers for interleaved practice—including perceptions that the strategy is more difficult, more time-consuming, or less interesting than

blocked practice. Indeed, interleaved practice was judged by more than 60% of students to be more difficult and more time-consuming than blocked practice—and, importantly, students are correct in these perceptions. Interleaved assignments demand switching between different problem types and identifying an appropriate approach for each one, which is more time-consuming and effortful compared to a blocked assignment that repeats one problem type over and over. These inherent attributes of interleaved practice (difficult, time-consuming) may discourage its use, especially if these attributes are interpreted as signs of low efficacy or generate dislike. Consistent with prior work, we found that higher perceived difficulty was associated with lower perceived efficacy ($r = -0.28$; see Table 1), highlighting that high effort is commonly misinterpreted as poor learning (Kirk-Johnson et al., 2019). Perceived difficulty was also negatively associated with low preference for interleaved (vs. blocked) practice ($r = -0.61$), suggesting that the difficulty of interleaved practice contributes to its low likability. Getting students to embrace the difficulty that is often inherent in effective learning strategies will be important not only for improving students' knowledge of strategy efficacy but also (and perhaps more) important for increasing overall strategy likability and preference.

4.1. Limitations

Our survey focused on 7th graders' perceptions of math learning strategies, so it is reasonable to ask whether our findings would generalize to students of other ages or to other subject areas. Students' strategy perceptions could change across grade levels as students acquire new learning experiences that shape their beliefs and expectations (Schneider, 2008). That said, even college students have difficulty recognizing the efficacy of spaced and interleaved math practice (Hartwig et al., 2022), so experience with school in general or math learning specifically might not, by itself, markedly change these strategy beliefs. Whether the same is true for strategy likability is unclear, though if likability is shaped by the perceived difficulty or efficiency of a strategy, we might predict continuity for likability as well. With respect to other subject areas, an interesting question is whether norms of practice within different fields of study might create different expectations and strategy beliefs among students within those fields. Notably, in the field of math (and other computational fields, such as chemistry or physics) students are commonly assigned practice problems, which serve as small units of practice that can be easily spaced across time or interleaved within assignments, providing many opportunities to encounter these learning strategies. Despite likely exposure to spaced and interleaved practice in the context of math learning, students still tend to undervalue these strategies. It may be useful for future work to verify whether spacing and interleaving are similarly undervalued by students in other fields, since these strategies are useful for other subject areas also (e.g., Carpenter et al., 2012; Eglington & Kang, 2017; Pan et al., 2019).

The present studies did not test whether the math strategy perceptions we surveyed did indeed affect actual use of spaced and interleaved practice, so we cannot say to what extent these perceptions mapped onto behavior. Rather, we rely on prior theorizing, models of self-regulated learning, and related work (e.g., McDaniel & Einstein, 2020; Obergriesser & Stoeger, 2020; Panadero, 2017; Zepeda et al., 2020) to underpin the assertion that students' strategy perceptions influence their strategy choices and that a strategy viewed as ineffective or unlikable will be used less often. Plausibly, students' strategy perceptions are less consequential when external forces (such as teachers' lesson plans or textbook assignments) dictate the type of practice that occurs. However, in many learning situations, students are decision-makers about the type of study they choose or the mental activities they engage—and their learning-related perceptions may become increasingly important as they get older and their learning activities become more self-regulated. Even when teachers

do heavily shape the practice that occurs, teachers themselves can be affected by students' perspectives and attitudes toward specific learning strategies, opting to encourage those strategies their students favor rather than protest. An interesting direction for further work may include teachers' perceptions of learning strategies (Halamish, 2018; Morehead et al., 2016; Rohrer et al., 2020b) and how teachers' choices are affected by their students' perceptions.

Also worth considering is whether participants could be justified in perceiving greater efficacy for strategy options we regarded as less efficacious. Though evidence from prior research is strong in support of spaced and interleaved practice, we did not verify the efficacy of the options shown in our scenarios. The first scenario in Study 1 may not always favor more (rather than less) practice, as prior work on overlearning has shown diminishing returns of consecutive repetitions of the same problem type within a single session (e.g., Rohrer & Taylor, 2006). Thus, the small proportion of students (15%) who perceived greater efficacy for fewer problems might have prioritized efficiency, and those who viewed the options as equally helpful (17%) may have correctly seen no additional benefit from extra problems. That said, the problems in that scenario (Figure 1) varied in form (sometimes requiring solving for area, other times solving for radius), so overlearning may be less relevant. With the interleaved scenario (Figure 3), participants could justifiably question the efficacy of the interleaved option if they noticed that topics taught later (e.g., on day 6) received less practice compared to earlier topics and compared to the same topics in the blocked option. This difference, typical of interleaving in classrooms, occurs because the first assignments must interleave with previously-taught topics (from a previous exam or previous school year) rather than with topics not-yet-taught, and later topics have fewer days available for interleaved practice before an exam (though they may be interleaved in future assignments beyond that exam). This non-equivalence between blocked and interleaved practice could have contributed to participants' perceptions of lower efficacy for interleaving, and future work might design scenarios with better equivalence. Regardless, effect sizes for interleaved (vs. blocked) practice are typically large and unlikely to be reversed by a small reduction of practice for later topics. Finally, misjudgments of efficacy could also result from misunderstanding the diagrams. For instance, when shown blocked and interleaved sample worksheets (Figure 5), did participants think that interleaving implied less practice or less instruction for each problem type? Future work may aim to clarify such details. In sum, we cannot say with certainty that participants' misperceptions of efficacy were entirely unjustified. However, our use of different diagrams across two studies offers some assurance that misperceptions were robust and not due to a single detail. As discussed above, our findings align with prior studies, using other methods, that also conclude that students underestimate the efficacy of interleaved practice.

4.2. Implications

Because students' perceptions of learning strategies can affect whether those strategies are used, it is important to understand how the most effective strategies are perceived by students. Interleaved practice, a highly effective strategy for math learning, may meet resistance because students dislike it and fail to recognize its efficacy. To promote interleaved practice, the challenge is, at minimum, twofold: convince learners of its efficacy and increase its palatability. To convince learners of its efficacy, approaches might include study strategy training or learning demonstrations, as well as explaining to students the value of the effort involved in interleaved practice. Approaches to improve the palatability of interleaved practice might include scaffolding, teamwork, or gamification (for more suggestions, see Zepeda et al., 2020).

In contrast, spaced practice may meet less resistance because students find it palatable, even if they are not always aware of the superior efficacy of more spaced (vs. less spaced) practice. To promote spacing, training and demonstrations may help teach its efficacy; but it is also possible that other barriers not examined here may pose a challenge for spacing. For example, when studying on their own, students may find it difficult to summon the discipline required to implement a continued schedule of spaced practice across time. Effective interventions will be those that help learners overcome the specific barriers that hinder the use of effective learning strategies. Importantly, as we show here, effective learning strategies do not all have the same barriers, and interventions should be tailored accordingly.

Author Contributions: Conceptualization, M.K.H. and D.R.; methodology, M.K.H. and D.R.; formal analysis, M.K.H.; data curation, M.K.H.; writing—original draft preparation, M.K.H.; writing—review and editing, M.K.H. and D.R.; visualization, M.K.H. and D.R.; project administration, M.K.H. and D.R.; funding acquisition, D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Institute of Education Sciences, U.S. Department of Education, grant R305A160263 to the University of South Florida. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Institutional Review Board Statement: This research was conducted in accordance with the Declaration of Helsinki. The Institutional Review Board of the University of South Florida determined the research to be exempt from full board review, as it was conducted in established or commonly accepted educational settings, involving normal educational practices, such as (i) research on regular and special education instructional strategies, or (ii) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original data presented in these studies are openly available at https://osf.io/rwbq4/ (accessed on 3 June 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

1 We note that the superiority of interleaved over blocked practice does not hold in all contexts. For example, blocking can outperform interleaving when learners must abstract and apply rules that define category membership, especially if within-category similarity is low (for further discussion, see Little & Nepangue, 2025; Noh et al., 2016). However, in classrooms, students are usually explicitly taught the rules that define to-be-learned concepts or categories rather than having to abstract them. Thus, for math learning in schools, which is the focus of this paper, we would expect the typical benefit of interleaved over blocked practice.

2 Judgments of preference (Study 2) were considered comparable to judgments of liking (Study 1), though we acknowledge that preference might be shaped by other factors in addition to liking, including familiarity or other justifications for favorability.

3 We inserted the name of a state standardized test that was familiar to the students in this study.

References

Barzagar Nazari, K., & Ebersbach, M. (2019). Distributing mathematical practice of third and seventh graders: Applicability of the spacing effect in the classroom. *Applied Cognitive Psychology*, *33*, 288–298. [CrossRef]

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444. [CrossRef]

Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: Comparing intended versus actual study behaviour. *Memory*, *25*, 784–792. [CrossRef]

Carpenter, S. K. (2014). Spacing and interleaving of study and practice. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying the science of learning in education: Infusing psychological science into the curriculum* (pp. 131–141). American Psychological Association.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*, 369–378. [CrossRef]

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, *1*, 496–511. [CrossRef]

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380. [CrossRef] [PubMed]

Cohen, M. S., Yan, V. X., Halamish, V., & Bjork, R. A. (2013). Do students think that difficult or valuable materials should be restudied sooner rather than later? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1682–1696. [CrossRef] [PubMed]

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation*, *53*, 63–147. [CrossRef]

Dunlosky, J. (2013). Strengthening the student toolbox: Study strategies to boost learning. *American Educator*, *37*(3), 12–21.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*, 4–58. [CrossRef]

Eglington, L. G., & Kang, S. H. K. (2017). Interleaved presentation benefits science category learning. *Journal of Applied Research in Memory and Cognition*, *6*, 475–485. [CrossRef]

Emeny, W. G., Hartwig, M. K., & Rohrer, D. (2021). Spaced mathematics practice improves test scores and reduces overconfidence. *Applied Cognitive Psychology*, *35*, 1082–1089. [CrossRef]

Firth, J., Rivers, I., & Boyle, J. (2021). A systematic review of interleaving as a concept learning strategy. *Review of Education*, *9*, 642–684. [CrossRef]

Foster, N. L., Mueller, M. L., Was, C., Rawson, K. A., & Dunlosky, J. (2019). Why does interleaving improve math learning? The contributions of discriminative contrast and distributed practice. *Memory & Cognition*, *47*, 1088–1101. [CrossRef]

Gay, L. R. (1973). Temporal position of reviews and its effect on the retention of mathematical rules. *Journal of Educational Psychology*, *64*, 171–182. [CrossRef]

Halamish, V. (2018). Pre-service and in-service teachers' metacognitive knowledge of learning strategies. *Frontiers in Psychology*, *9*, 2152. [CrossRef]

Hartwig, M. K., Rohrer, D., & Dedrick, R. F. (2022). Scheduling math practice: Students' underappreciation of spacing and interleaving. *Journal of Experimental Psychology: Applied*, *28*, 100–113. [CrossRef] [PubMed]

Hopkins, R. F., Lyle, K. B., Hieb, J. L., & Ralston, P. A. (2016). Spaced retrieval practice increases college students' short- and long-term retention of mathematics knowledge. *Educational Psychology Review*, *28*, 853–873. [CrossRef]

Kang, S. H. (2016). Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences*, *3*, 12–19. [CrossRef]

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, *115*, 101237. [CrossRef]

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*, 585–592. [CrossRef]

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, *25*, 498–503. [CrossRef] [PubMed]

Little, J. L., & Nepangue, J. A. (2025). Whether interleaving or blocking is more effective for long-term learning depends on one's learning strategy. *Behavioral Sciences*, *15*, 662. [CrossRef] [PubMed]

Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. S. (2020). How the amount and spacing of retrieval practice affect the short- and long-term retention of mathematics knowledge. *Educational Psychology Review*, *32*, 277–295. [CrossRef]

Mayfield, K. H., & Chase, P. N. (2002). The effects of cumulative practice on mathematics problem solving. *Journal of Applied Behavior Analysis*, *35*, 105–123. [CrossRef]

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*, 462–476. [CrossRef]

McDaniel, M. A., & Einstein, G. O. (2020). Training learning strategies to promote self-regulation and transfer: The knowledge, belief, commitment, and planning framework. *Perspectives on Psychological Science*, *15*, 1363–1381. [CrossRef]

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*, 257–271. [CrossRef]

Noh, S. M., Yan, V. X., Bjork, R. A., & Maddox, W. T. (2016). Optimal sequencing during category learning: Testing a dual-learning systems perspective. *Cognition*, *155*, 23–29. [CrossRef]

Obergriesser, S., & Stoeger, H. (2020). Students' emotions of enjoyment and boredom and their use of cognitive learning strategies—How do they affect one another? *Learning and Instruction*, *66*, 101285. [CrossRef]

Pan, S. C. (2015). The interleaving effect: Mixing it up boosts learning. *Scientific American Mind*, *313*(2). Available online: https://www.scientificamerican.com/article/the-interleaving-effect-mixing-it-up-boosts-learning/ (accessed on 29 July 2025).

Pan, S. C., Tajran, J., Lovelett, J., Osuna, J., & Rickard, T. C. (2019). Does interleaved practice enhance foreign language learning? The effects of training schedule on Spanish verb conjugation skills. *Journal of Educational Psychology*, *111*, 1172–1188. [CrossRef]

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422. [CrossRef]

Rau, M. A., Aleven, V., & Rummel, N. (2013). Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave? *Learning and Instruction*, *23*, 98–114. [CrossRef]

Rea, S. D., Wang, L., Muenks, K., & Yan, V. X. (2022). Students can (mostly) recognize effective learning, so why do they not do it? *Journal of Intelligence*, *10*, 127. [CrossRef]

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition*, *1*, 242–248. [CrossRef]

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review*, *24*, 355–367. [CrossRef]

Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review 27*, 635–643. [CrossRef]

Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review*, *21*, 1323–1330. [CrossRef]

Rohrer, D., Dedrick, R. F., & Hartwig, M. K. (2020a). The scarcity of interleaved practice in mathematics textbooks. *Educational Psychology Review*, *32*, 873–883. [CrossRef]

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C.-N. (2020b). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology*, *112*, 40–52. [CrossRef]

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, *107*, 900–908. [CrossRef]

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology*, *20*, 1209–1224. [CrossRef]

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics practice problems boosts learning. *Instructional Science*, *35*, 481–498. [CrossRef]

Sana, F., Yan, V. X., & Kim, J. A. (2017). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology*, *109*, 84–98. [CrossRef]

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education*, *2*, 114–121. [CrossRef]

Schutte, G. M., Duhon, G. J., Solomon, B. G., Poncy, B. C., Moore, K., & Story, B. (2015). A comparative analysis of massed vs. distributed practice on basic math fact fluency growth rates. *Journal of School Psychology*, *53*, 149–159. [CrossRef] [PubMed]

Son, L. K., & Simon, D. A. (2012). Distributed learning: Data, metacognition, and educational implications. *Educational Psychology Review*, *24*, 379–399. [CrossRef]

Susser, J. A., & McCabe, J. (2013). From the lab to the dorm room: Metacognitive awareness and use of spaced study. *Instructional Science*, *41*, 345–363. [CrossRef]

Taraban, R., Maki, W. S., & Rynearson, K. (1999). Measuring study time distributions: Implications for designing computer-based courses. *Behavioral Research Methods, Instruments, and Computers*, *31*, 263–269. [CrossRef]

Taylor, K., & Rohrer, D. (2010). The effect of interleaving practice. *Applied Cognitive Psychology*, *24*, 837–848. [CrossRef]

Weinstein, Y., Madan, C. R., & Sumeracki, M. A. (2018). Teaching the science of learning. *Cognitive Research: Principles and Implications*, *3*, 2. [CrossRef]

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*, 568–579. [CrossRef]

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*, 918–933. [CrossRef]

Yan, V. X., & Sana, F. (2021). Does the interleaving effect extend to unrelated concepts? Learners' beliefs versus empirical evidence. *Journal of Educational Psychology*, *113*, 125–137. [CrossRef]

Yan, V. X., Soderstrom, N. C., Seneviratna, G. S., Bjork, E. L., & Bjork, R. A. (2017). How should exemplars be sequenced in inductive learning? Empirical evidence versus learners' opinions. *Journal of Experimental Psychology: Applied*, *23*, 403–416. [CrossRef] [PubMed]

Zepeda, C. D., Martin, R. S., & Butler, A. C. (2020). Motivational strategies to engage learners in desirable difficulties. *Journal of Applied Research in Memory and Cognition*, *9*, 468–474. [CrossRef]

Ziegler, E., & Stern, E. (2014). Delayed benefits of learning elementary algebraic transformations through contrasted comparisons. *Learning and Instruction*, *33*, 131–146. [CrossRef]

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Students', Teachers', and Parents' Knowledge About and Perceptions of Learning Strategies

Amber E. Witherby [1,*], Addison L. Babineau [2] and Sarah K. Tauber [3]

[1] Department of Psychological Sciences, Creighton University, Omaha, NE 68178, USA
[2] Department of Psychology & Philosophy, Texas Woman's University, Denton, TX 76204, USA; ababineau@twu.edu
[3] Department of Psychology, Texas Christian University, Fort Worth, TX 76109, USA; uma.tauber@tcu.edu
* Correspondence: amberwitherby@creighton.edu

Abstract

Previous research has demonstrated that students have imperfect knowledge about learning strategies. Moreover, very little is known about teachers' and parents' knowledge about learning strategies, which is important as these groups can help to model and teach students effective strategies. The goal of the present research was to add to this literature by measuring students', teachers', and parents' beliefs about several learning strategies that have received empirical scrutiny, using methodology that builds upon prior work. To do so, participants were presented with a series of learning scenarios exemplifying a different learning strategy and rated each individually on effectiveness, familiarity, and their likelihood of using (or recommending) it in the future. Participants had accurate knowledge about effective strategies, rating retrieval practice and distributed study as the most effective learning strategies. There were variations within groups on their levels of familiarity with each strategy, which may have informed their ratings of effectiveness. For instance, participants rated interleaving as less effective compared to blocking but were also less familiar with the former. On a positive note, these outcomes suggest that people generally have good knowledge about learning strategies and underscore the importance of a broad dissemination of educational research.

Keywords: learning strategies; metacognitive knowledge; teachers; parents; survey

1. Introduction

Metacognition is broadly defined as a person's thoughts about cognitive phenomena (Flavell, 1971, 1979). Key components of metacognition include metacognitive knowledge (i.e., people's knowledge about cognition), metacognitive monitoring (i.e., people's assessments of cognitive performance), and metacognitive control (i.e., decisions people make about how to approach a cognitive task) (for reviews, see Dunlosky & Metcalfe, 2009; Tauber & Dunlosky, 2016). Most important for the present purposes, having accurate metacognitive knowledge is an important pre-cursor for students to use effective learning strategies. Although ample research has identified highly effective learning strategies, it is unclear how knowledgeable people are on this topic (Dunlosky et al., 2013; Fiorella & Mayer, 2016; Miyatsu et al., 2018). Accordingly, we administered a survey to college students to evaluate their knowledge about and perceptions of several learning strategies. As well, given that people are not born with knowledge about how to learn, it is important to understand how (and from where) people develop this knowledge. For students,

teachers and parents likely serve as the primary resource for learning how to learn. As such, it is important to evaluate whether teachers and parents have accurate knowledge about learning strategies. To that end, we also evaluated teachers' and parents' knowledge and perceptions.

1.1. Which Strategies Are Effective for Learning?

To contextualize people's knowledge about learning strategies, it is necessary to start with a brief review of the literature evaluating the effectiveness of these strategies. In general, strategies that are more passive on the students' part (e.g., rereading, highlighting) tend to be less effective compared to strategies that require active engagement with the content (e.g., retrieval practice, elaborative interrogation) (Dunlosky et al., 2013; Fiorella & Mayer, 2016). Moreover, the effectiveness of learning strategies can vary depending on characteristics of the to-be-learned material and of the learner. To attempt to quantify the effectiveness of commonly investigated learning strategies, Dunlosky and colleagues reviewed the literature for 10 learning strategies, assigning each a "utility rating" (Dunlosky et al., 2013). Two strategies, retrieval practice and distributed study, received high utility ratings, indicating that their effectiveness is robust across various contexts and learners. Four strategies (blocking content order, interleaving content order, elaborative interrogation, and self-explanation) received moderate utility ratings, suggesting that they can be effective for some types of materials. Finally, six strategies (highlighting, rereading, keyword mnemonic, imagery, summarization, using a massed restudy schedule) received low utility ratings, suggesting that their effectiveness may be limited to certain contexts and/or learners.

From this review, retrieval practice and using a distributed study schedule stand above other strategies with regard to their general effectiveness. However, it becomes more challenging to draw clear conclusions about whether one strategy is more effective than another within each level of utility rating (e.g., summarization versus elaborative interrogation). Moreover, although these are the general conclusions about the effectiveness of these strategies, it is important to note that the way that students implement a strategy can moderate its effectiveness. As one example, rereading can be more effective if used in combination with a distributed restudy schedule as compared to a massed restudy schedule (Miyatsu et al., 2018). Likewise, highlighting and underlining can be effective learning strategies if students are strategic with them and are able to identify information that is important (Miyatsu et al., 2018). As a final example, the keyword mnemonic can be particularly helpful for learning various types of paired associates (e.g., foreign language vocabulary and first language vocabulary; for a review, see Pressley et al., 1982; U.S. states and their capitals; Levin et al., 1980). Thus, when measuring people's beliefs about learning strategies, it is important to ensure they all consider the same form of implementation when evaluating a strategy's effectiveness. In addition, the effectiveness of a strategy may depend on students' ability. For instance, the effectiveness of summarization is moderated by the quality of students' summarizations (Fiorella & Mayer, 2016). That is, students who produce high-quality summarizations will retain more compared to students who produce low-quality summarizations. Thus, although some strategies generally appear more effective than others (e.g., retrieval practice relative to rereading), we encourage readers to consider these strategies flexibly in the context of when and how they are being used.

1.2. Students' Knowledge and Use of Learning Strategies

Several researchers have explored the frequency with which students report using various learning strategies (Anthenien et al., 2018; Blasiman et al., 2017; Geller et al.,

2018; Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007; Morehead et al., 2016; Yan et al., 2014; Zepeda & Nokes-Malach, 2021). A consistent outcome from these studies is that students use a variety of learning strategies, many of which are suboptimal. For instance, as reviewed above, two of the most effective strategies for enhancing learning are distributed study (i.e., spacing study sessions out over time) and retrieval practice (i.e., self-testing). When asked how students space their study, about 52% reported cramming their studying into a single session, which is the less effective strategy compared to spacing study over multiple sessions (Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007; Morehead et al., 2016). It is unclear why so many students report using the less effective strategy. One possibility is that students have incomplete knowledge about learning. Consistent with this possibility, of the roughly 70% of students who reported using retrieval practice, most reported using it to monitor their learning rather than to help them learn the information. Thus, students appear to have incomplete knowledge about why self-testing is beneficial. Although retrieval practice can be beneficial for helping students identify what they do and do not know, it is also beneficial in that the act of retrieving information from memory increases the memory strength of that information (e.g., Roediger & Butler, 2011).

Although these studies are informative for understanding how students study, they do not provide clear insight into students' knowledge about the effectiveness of learning strategies. That is, just because a student uses (or does not use) a strategy, it does not necessarily mean that they think that strategy is effective (or ineffective). Rather, many things can influence whether students use a particular learning strategy (e.g., the amount of time available, the amount of effort they are willing to invest) (Rea et al., 2022). To evaluate students' knowledge about the effectiveness of learning strategies, McCabe (2011) presented students with a series of learning scenarios that pitted learning strategies against each other (e.g., self-testing vs. restudying; spacing vs. massing study) and had them choose the strategy that would be most beneficial for learning. Students generally did not endorse the empirically supported strategy. For instance, students rated massing as more effective than spacing and rereading as more effective than self-testing.

As reviewed above, students appear to hold inaccurate knowledge about the effectiveness of common learning strategies when they make relative comparisons; however, students' knowledge about individual learning strategies cannot be determined in a paradigm that pits strategies against each other. For instance, students may believe that self-testing and rereading are both effective, but they may rate rereading as somewhat more effective if forced to choose between them. To evaluate students' knowledge about learning strategies in isolation, Blasiman et al. (2017) had students rate the effectiveness of 10 common learning strategies. Students rated reading notes as the most effective strategy and self-testing as one of the least effective strategies. Relatedly, Hartwig et al. (2022) found that students underappreciated the benefits of spaced study schedules and interleaving content order (i.e., mixing problems of different types) when evaluating the utility of these strategies. Similarly, even when students are given experience with more effective (i.e., interleaved study schedule) and less effective (i.e., blocking study schedule) learning strategies, students still report a preference for the less effective strategy (Yan et al., 2016). Thus, students' knowledge about the effectiveness of these learning strategies was far from perfect.

Why might students have inaccurate knowledge about learning strategies? One possibility is that they were never taught about effective learning techniques by their teachers or parents. Indeed, Morehead et al. (2016) found that only 36% of students reported using strategies that they learned from their teacher. This low percentage is surprising given that, in surveys of college teachers, most reported discussing learning strategies in class (79%, Morehead et al., 2016; 91%, Hunter & Lloyd, 2018). Even so, if students learn about strategies from their teachers (or parents), it is necessary to know

whether these groups have accurate knowledge about learning strategies. If not, they may not provide effective advice.

1.3. Teachers' and Parents' Knowledge About Learning Strategies

Few studies have explored teachers' knowledge about learning strategies. The available literature suggests that teachers' knowledge about learning strategies is also imperfect (Firth, 2021; Halamish, 2018; Wissman & Leontyev, 2024). Firth (2021) and Halamish (2018) presented high school teachers with learning scenarios like those used by McCabe (2011) in which two learning strategies were pitted against each other, and the teachers predicted which strategy was more effective for enhancing learning. The teachers generally did not endorse the empirically supported learning strategies. For instance, only about 30% of the teachers endorsed retrieval practice and 50% of the teachers endorsed spacing study (Firth, 2021). Using similar methodology with college teachers, Morehead et al. (2016) found somewhat more promising outcomes. Specifically, 62% of the teachers endorsed retrieval practice and 74% endorsed spacing study. In addition, the teachers reported recommending a combination of effective strategies (e.g., retrieval practice) and less effective strategies (e.g., rereading).

Parents can also serve as an important resource for students' development of learning skills (Lieder et al., 2023). In a survey with high school students, 35% of students on a high-achievement track (8% of students on a low-achievement track) reported seeking academic assistance from their parents (Zimmerman & Pons, 1986). Hoover-Dempsey et al. (1995) interviewed parents of children in 1st–5th grade about their involvement with their children's homework. Most parents (97%) reported helping their children in some capacity, but many of them reported feeling unprepared. In addition, only about half of the parents (55%) reported discussing learning strategies with their children. There is evidence that parental involvement with their child's learning is positively related to that child's academic achievement (Cheung & Pomerantz, 2012; Martinez-Pons, 2002). This may be because parents who are involved teach their children effective learning strategies.

Consistent with this possibility, Lieder et al. (2023) found that parents' beliefs about strategies for learning math predicted students' use of those strategies about a year later. In their study, they surveyed parents and children (13–15 years old) at two time points about a year apart. Parents and children rated the effectiveness of 21 strategies that could be used to learn math (e.g., *I work on math problems in textbooks; I create my own math questions; when I study mathematics, I try to understand the relationship between different contents/topics*). As well, children also rated how frequently they used each strategy. They found that parents' ratings of effectiveness at Time 1 predicted students' use of the strategy at Time 2. Moreover, students' ratings of effectiveness at Time 2 mediated this relationship, suggesting that parents' beliefs impacted students' beliefs, which in turn influenced their learning strategy use. Although this study highlights the role that parents' knowledge can play in students' knowledge of learning strategies, this study was limited to parents and children in Japan and to one content domain (math). Moreover, this study only explored whether parents' beliefs were related to students' beliefs and not whether parents have accurate knowledge about effective learning strategies. Critically, no research has explored the latter issue. Thus, it is unclear whether parents have accurate knowledge about learning strategies.

1.4. The Present Study

In sum, several questions remain about the knowledge that students, teachers, and parents hold about learning strategies. The primary motivation of the present study was to address these questions. Specifically, we aimed to address the limitations and gaps in prior work in the following ways:

1. *Measuring participants knowledge about learning strategies without pitting strategies against each other*. By forcing participants to pick one of two strategies, it is challenging to draw conclusions about their knowledge. For instance, participants may believe both provided strategies are effective but artificially choose one over the other. In the present study, we allowed participants to rate each strategy in isolation, which allows for a clearer interpretation of their knowledge about that strategy.
2. *Providing concrete examples about how a strategy is implemented*. Providing concrete examples ensures that participants consider each strategy similarly. As reviewed above, the way strategies are implemented can moderate their effectiveness (Dunlosky et al., 2013; Miyatsu et al., 2018). Thus, to accurately measure and compare participants' ratings of a strategy, participants need to consider the same method of implementation. Moreover, providing concrete examples may be especially important if people are unfamiliar with a strategy. Thus, in the present study, we provided participants with a concrete example of one way the strategy could be implemented.
3. *Increasing data on teachers' knowledge about learning strategies*. As reviewed above, few studies have investigated teachers' knowledge about learning strategies. Moreover, those that did included the limitations described in points 1 and 2[1]. Thus, the present study aimed to provide more data on teachers' knowledge while addressing these issues.
4. *Gathering novel data on parents' knowledge about learning strategies*. No research has directly evaluated parents' knowledge of common learning strategies. Rather, researchers have focused on parental involvement with homework and how parents' beliefs about learning strategies in math are related to students' use of those strategies. In the present study, we directly evaluated parents' knowledge of the effectiveness of common learning strategies with a sample of parents with somewhat older children than has been investigated previously (i.e., middle and high school children).

We also included several additional novel features to build upon prior work. We investigated learning strategies that had not been previously investigated (e.g., interleaved restudy schedule, elaborative interrogation). In addition to rating the effectiveness of each strategy, participants rated their familiarity with it and their likelihood of using (students) or recommending (teachers and parents) it. These novel ratings are useful for providing insight into participants' ratings of effectiveness. For instance, if a strategy is given a low effectiveness rating, it could be because participants are unaware of it (e.g., a low familiarity rating) or because they know about the strategy and believe that it is ineffective (e.g., a high familiarity rating). Additionally, students may have accurate knowledge about learning strategies, yet they still may choose not to use effective strategies because of various barriers such as time cost or low self-efficacy (Rea et al., 2022).

2. Materials and Methods

2.1. Participants

We used past research on surveys evaluating the effectiveness of learning strategies as a guide for our target sample size (Blasiman et al., 2017; Morehead et al., 2016). Our target goal was to recruit 150 participants per group. Consistent with prior research, we collected our student sample from an undergraduate university research participation pool. To obtain a representative and diverse sample for our parent and teacher samples, we recruited participants from multiple sources (see below).

A total of 151 undergraduate students were recruited from the Texas Christian University (TCU) psychology research pool and participated in exchange for partial course credit. Students were college-aged (M = 20.23, SE = 0.29) and predominantly identified as a woman and White (see Table 1 for more detailed demographic information). Students

had completed an average of 2 years of college (i.e., total years of education, M = 13.94, SE = 0.11) and reported an average college GPA of 3.4 (SE = 0.04). Most students (84.1%, n = 127) reported that they were not first-generation college students (i.e., most students had at least one parent with a college degree).

Table 1. Demographic Information.

Question	Response Options	Students	Parents	Teachers
What is your gender?	Male	21.2	32.2	36.1
	Female	78.8	67.1	61.8
	Prefer not to respond	0.0	0.7	2.1
What is your ethnicity?	Caucasian/White (non-Hispanic)	68.2	77.6	74.3
	Black (non-Hispanic)	4.6	5.6	2.6
	Asian or Pacific Islander	9.3	4.9	5.8
	American Indian	0.7	0.7	0.0
	Hispanic	15.9	7.7	11.5
	Other	1.3	3.5	2.1
	Prefer not to respond	0.0	0.0	3.7
What is your highest level of education? [1]	Some High School	-	2.1	0.0
	Completed High School	-	11.2	0.0
	Obtained GED	-	2.8	0.0
	Some College	-	25.9	0.0
	Associate's Degree	-	14.7	0.0
	Bachelor's Degree	-	30.7	20.9
	Master's Degree	-	11.9	44.5
	Doctoral Degree	-	0.7	34.0
	Prefer not to respond	-	0.0	0.5

Note: Value represents the percent of the sample that selected each option. [1] This question was only asked to parents and teachers.

A total of 169 parents completed the survey and were paid $10 (USD) for their participation. Parents were recruited from a TCU community forum (n = 11), ads posted on Facebook (n = 12), Amazon's Mechanical Turk (mTurk) Cloud Research (n = 118)[2], and a laboratory database (n = 28).[3] We had four inclusion criteria for the parent sample: (a) have a child aged 10 to 17 years old, (b) have no current or prior experience as a teacher, (c) speak English, and (d) live in the United States. Of the 169 parents, all responses from 26 (15.4%) were excluded because the participant indicated they had current or prior experience teaching (laboratory database, n = 13; mTurk, n = 9; TCU community forum, n = 4)[4]. Thus, the final parent sample consisted of 143 participants. Parents were middle-aged (M = 42.50, SE = 1.91) and predominantly identified as a woman and White (see Table 1). Most parents reported having at least some college experience (83.9%). Parents also reported demographic information about their child. The average age of the parents' children was 13.54 years old (SE = 0.19), and the majority were identified as a boy (67.8%, n = 97). The children ranged from being in 4th–12th grade.

A total of 194 teachers were recruited and were paid $10 (USD) for their participation. Teachers were recruited via targeted e-mails[5] (n = 99), mTurk Cloud Research (n = 44), ads posted on Facebook (n = 28), from a local middle school (n = 15), and via referral from other teachers (n = 8). We had three inclusion criteria for the teacher sample: (a) they had to have experience working as a teacher at a middle school, high school, or college, (b) speak English, and (c) live in the United States. Prior to completing the survey, teachers were required to indicate the school at which they taught and the grade level(s) they taught. Of the 194 teachers, all responses from 3 (1.5%) were excluded because the participant did not include any information about their teaching experience (Facebook ad, n = 2; referral n = 1). Thus, the final teacher sample consisted of 191 participants. The average age for the teacher sample was 40.04 (SE = 0.82), and they predominantly identified as a woman and White

(see Table 1). All teachers held an advanced degree. On average, teachers in the sample had been teaching for more than a decade (M = 11.97 years, SE = 0.63). The teachers taught at a range of levels (college or university, n = 87; high school, n = 46, middle school, n = 54; multiple levels, n = 4) and taught a variety of topics including science (n = 59; e.g., biology, engineering, psychology, computer science), math (n = 53; e.g., algebra, statistics, calculus, geometry), English (n = 20; e.g., language arts, English as a second language, literature), social studies or liberal arts (n = 20; e.g., history, anthropology, government), as well as others (n = 52; e.g., foreign language, arts, business)[6]. The Texas Christian University Institutional Review Board approved this study.

2.2. Materials and Procedure

The survey was nearly identical for all groups. A copy of the questions that each group received can be found in the online supplemental materials "https://osf.io/ghtyn/". Participants were presented with a series of 12 learning scenarios that each described a hypothetical student using one of twelve learning strategies (see Table 2). The learning strategies were selected based on common strategies that students report using and that have received empirical scrutiny (for a review, see Dunlosky et al., 2013).

Table 2. Learning Scenarios used in the Present Study.

Learning Strategy	Learning Scenario
Elaborative Interrogation	One way to study is to come up with an explanation for something you are trying to learn. You can do this by asking yourself "why" questions. For example, when a student is trying to learn the second amendment of the constitution (i.e., right to bear arms), he may ask himself why this amendment was originally implemented.
Self-explanation	One way to study is to explain how you are thinking to yourself. For example, when a student is trying to learn about soundwaves, she may ask herself, "What do soundwaves mean to me?" or "What do I already know about soundwaves?".
Summarization	One way to study is to summarize the material you are learning as you go. For example, when a student is trying to learn information from his textbook (e.g., the events leading up to World War II), he may read a section of the text, and then write a 2–3 sentence summary of the main points of that section.
Highlighting	One way to study is to highlight (or underline) information that is perceived as being important in the text or in one's notes. For example, when a student is trying to learn about important historical figures, she may highlight (or underline) this information in her textbook while reading.
Keyword Mnemonic	One way to study is to generate keywords that link together concepts you are trying to learn. For example, perhaps a student is trying to learn Spanish vocabulary words, including the word vaca (which means cow). He may think of the word vacation in English because it sounds like vaca. Then, he thinks of an image of the two words interacting. For example, the student may think of a cow that is on vacation at a beach.
Imagery	One way to study is to create mental images as you read. For example, when a student is reading her textbook and trying to learn about the planets in the solar system, she may try to visualize all of the planets including their order relative to the sun, their color, and their size.
Rereading	One way to study is by rereading or restudying material. For example, when a student is trying to learn geography of the United States, he may study a map of the United States with all of the states labeled, and then revisit this map multiple times to study it. As another example, he may try to learn information about when each state was founded by rereading his textbook over and over.
Retrieval Practice	One way to study is by retrieving information from memory, without the answer present (i.e., practice testing). For example, when a student is trying to learn the phases of mitosis (a type of cell division), she may study by making flashcards for each phase, and then testing herself using the flashcards.
Distributed Restudy Schedule	One way to study is by spreading your study sessions out over time. For instance, if a student is learning new vocabulary words, he may choose to study them for 1 h per day, for 5 consecutive days.
Massed Restudy Schedule	One way to study is to study everything in one block of time. For instance, when learning about the different branches of government, a student may choose to study all three branches (legislative, executive, judicial) in one session for 5 h.

Table 2. *Cont.*

Learning Strategy	Learning Scenario
Interleaved Content Order	Imagine that a student has two tests next week: one in science and one in history. To study for these tests, the student may make a set of flashcards for his science test and a second set of flashcards for his history test. One way to study for these tests would be to intermix the flashcards from the two classes. That is, the student would alternate between studying flashcards for his science class and flashcards for his history class.
Blocked Content Order	Imagine that a student is trying to learn how to play basketball, which involves learning a number of different skills. One way to learn these skills is by practicing one skill at a time until it is mastered before moving on to the next skill. For example, a student may start by learning how to make a bounce pass. To do so, she would make bounce passes over and over until she has mastered that skill. Then, she may move on to learning the next skill, how to make a chest pass. She would then make chest passes over and over until she masters that skill. She would move on to the next skill (e.g., shooting free throws).

Before beginning the survey, participants were told that we were interested in understanding their beliefs about the effectiveness of various learning strategies. Specifically, students received the following instructions:

> "People have different ideas about what strategies are beneficial for student learning. Given that you (as a student) are often faced with the challenging task of learning information for class, we are interested in your beliefs about which strategies are beneficial for learning. In the following survey, we will give you examples of different learning strategies and ask you to rate them."

Teachers received these instructions:

> "People have different ideas about what strategies are beneficial for student learning. Given that you (as a teacher) often serve as a primary resource for your students when completing homework or studying for exams, we are interested in your beliefs about which strategies are beneficial for learning. In the following survey, we will give you examples of different learning strategies and ask you to rate them."

Finally, parents received these instructions:

> "People have different ideas about what strategies are beneficial for student learning. Given that you (as a parent) often serve as a primary resource for your child when completing homework or studying for exams, we are interested in your beliefs about which strategies are beneficial for learning. In the following survey, we will give you examples of different learning strategies and ask you to rate them."

The survey was entirely self-paced. After reading about each strategy, participants were asked to rate the effectiveness of the strategy on a scale from 1 (*very ineffective*) to 10 (*very effective*) and their familiarity with the strategy on a scale from 1 (*very unfamiliar*) to 10 (*very familiar*). Students were asked how likely they would be to use the strategy, whereas parents and teachers were asked how likely they would be to recommend the strategy to their child or student on a scale of 1 (*very unlikely*) to 10 (*very likely*). For each scenario, participants made the ratings in this order (i.e., effectiveness, familiarity, and the likelihood of use or recommendation). Ratings were made by clicking on a value from 1 to 10 (for participants who completed the survey electronically) or by circling a value from 1 to 10 (for participants who completed a physical copy of the survey).

As an additional novel component of our study, given that researchers have recently investigated whether the effectiveness of retrieval practice is moderated by retrieval format, we explored whether students' and teachers' knowledge about retrieval practice was influenced by retrieval format (i.e., overt vs. covert; Tauber et al., 2018). As such, students

and teachers were presented with an additional two learning scenarios (see Table 3). One scenario described overt retrieval (i.e., retrieval practice by physically recalling an item), whereas the other described covert retrieval (i.e., retrieval practice by mentally recalling an item).

Table 3. Overt Retrieval and Covert Retrieval Scenarios used in the Present Study.

Learning Strategy	Learning Scenario
Overt Retrieval Practice	When studying, a student may decide to test himself using flashcards. For example, he may put a key term on one side of the flash card and a definition on the other. One way he could test himself is by **overtly** recalling the definition for each key term. That is, when he sees a new key term, he could **say the definition out loud** before checking the definition on the other side of the flash card.
Covert Retrieval Practice	When studying, a student may decide to test herself using flashcards. For example, she may put a key term on one side of the flash card and a definition on the other. One way she could test herself is by **covertly** recalling the definition for each key term. That is, when she sees a new key term, she could recall the definition **silently in her head** before checking the definition on the other side of the flash card.

Note. Bold and underlined formatting reflects how these scenarios were presented to participants.

Most participants (i.e., all students and parents and 92.3% of teachers) completed the survey on Qualtrics. All students and 15 parents completed the survey in the lab. The remaining parents and teachers completed the survey on their own time from their own personal computers. For these participants (i.e., all of whom completed the survey on Qualtrics), the order of the learning scenarios was quasi-randomized in a new order for each participant. Specifically, the order of the 12 primary learning strategies was randomized anew for each participant with the constraint that, for students and teachers, the overt retrieval scenario and the covert retrieval scenario immediately followed the retrieval practice scenario. The order of the overt and covert scenarios was randomized anew for each participant.

Fifteen teachers (i.e., 7.7%) completed the survey using a physical paper copy, which they completed on their own time and mailed back to the researchers. The order of the learning scenarios was quasi-randomized. Specifically, participants completed the survey in the same random order, with the constraint that the overt retrieval scenario and the covert retrieval scenario immediately followed the retrieval practice scenario. After participants rated the retrieval practice scenario, they rated the overt retrieval practice scenario, followed by the covert retrieval practice scenario.[7]

Students and teachers were given an additional six multiple-choice questions to evaluate their beliefs about notetaking and learning styles. These questions were adapted from Witherby and Tauber (2019). Specifically, students were asked how frequently (on a 5-point scale from *never* to *always*) their professors (a) use PowerPoint to present information to their class, (b) give them instructions regarding how they should take notes, and (c) give them handouts to facilitate their notetaking. Teachers received these same questions, except they were framed around how often they did each of these things (e.g., *how often do you give your students instruction regarding how they should take notes?*). Students were also asked to estimate the percentage of information from class that they record in their notes (teachers were asked to estimate the percentage of information from class that they thought students recorded in their notes). Students and teachers responded to this question by selecting one of five options: 0–20%, 21–40%, 41–60%, 61–80%, 81–100%. The order of the notetaking questions was randomized for participants who completed the survey on Qualtrics and fixed in the order described for teachers who completed the survey with a physical copy. Next, students and teachers were asked whether they believe students have different learning styles (e.g., visual vs. auditory learners) and responded by selecting “yes” or “no”.

Finally, students were asked whether they think professors should teach in a way that accommodates those differences (teachers were asked whether they teach in a way that accommodates those differences) and responded by selecting, "yes", "no", or "I did not say yes". The response options for these six multiple-choice questions were presented in a fixed order.[8]

Finally, all participants answered a series of demographic questions (e.g., age, gender, ethnicity). Teachers were asked questions about their teaching (e.g., how many years they have been teaching, what grade levels they teach, what topics they teach). In addition, students and teachers were asked to rate how confident they are that they know which strategies are effective for learning and responded by selecting one of five options: 0–20%, 21–40%, 41–60%, 61–80%, 81–100%. Students and teachers were also asked whether they have previously received special training on student learning strategies and responded by selecting "yes" or "no". If they selected yes, they were then given a free-response question to explain. Students and teachers were asked whether they would attend a workshop to receive training on student learning strategies and responded by selecting "yes" or "no". Teachers were asked whether they think it would be important for teachers to attend a workshop in which they receive training on student learning strategies and responded on a 5-point scale from "*strongly disagree*" to "*strongly agree*". The order of the demographic questions (and these follow-up questions for students and teachers) was fixed in the order described.

3. Results

As reviewed above, ample evidence supports the conclusion that retrieval practice is a highly effective strategy for all types of learners to learn a wide variety of content (Dunlosky et al., 2013). As such, throughout the results, we use retrieval practice as a baseline effective strategy to which we will compare all other strategies.[9] To organize the results, we conducted a series of 2 (strategy) $\times$ 3 (group) mixed ANOVAs to evaluate (a) participants' beliefs for all strategies compared to retrieval practice and (b) whether these beliefs differed based on the sample (students, parents, teachers). Given that some strategies have a natural comparison strategy (e.g., distributed restudy schedule vs. massed restudy schedule), we also conducted 2 $\times$ 3 ANOVAs for these comparisons. Finally, given that only students and teachers answered questions about covert and overt retrieval practice, we conducted a 2 $\times$ 2 ANOVA for these comparisons. For all analyses, to follow up on significant main effects and interactions, we used the Bonferroni correction to account for multiple comparisons. As well, we checked for the assumption of normality using Levene's test for homogeneity of variance. When this assumption was violated, we reported follow-up comparisons using a Welch's ANOVA, which tests if the effect remains significant while accounting for unequal variances.

For interested readers, we have also conducted additional analyses that can be located in the online supplemental materials (https://osf.io/ghtyn/). Specifically, we conducted a series of one-way ANOVAs (12 levels) for each strategy separately for each measure (effectiveness, familiarity, and future use/recommendation) for each group (students, parents, and teachers). These outcomes can be found in Appendix B in the online supplemental materials. Additionally, we have also conducted a series of one-way ANOVAs (three levels) comparing each group separately for each strategy and measure. These outcomes can be found in Appendix C in the online supplemental materials.

3.1. Ratings of Effectiveness

In general, participants appeared to have accurate knowledge about two of the most effective strategies for improving student learning: using a distributed restudy schedule

and retrieval practice (see Table 4). Across groups, participants consistently gave the highest effectiveness ratings to these strategies. Participants in all groups also consistently gave lower ratings to interleaving content order and using a massed restudy schedule. The remaining strategies all generally fell between these ends, with variability between groups in how effective participants believed they were.

Table 4. Ratings of Effectiveness.

Learning Strategy	Students (n = 151)	Parents (n = 143)	Teachers (n = 191)
Highlighting	6.80 (0.21)	8.06 (0.16)	6.72 (0.18)
Rereading	7.43 (0.17)	7.64 (0.17)	6.59 (0.18)
Keyword Mnemonic	8.09 (0.16)	7.20 (0.20)	7.28 (0.15)
Imagery	7.74 (0.15)	8.05 (0.14)	7.88 (0.13)
Summarization	8.14 (0.14)	7.94 (0.16)	8.28 (0.11)
Massed Restudy Schedule	4.75 (0.19)	5.01 (0.21)	4.06 (0.17)
Blocked Content Order	7.38 (0.18)	7.71 (0.17)	7.35 (0.15)
Interleaved Content Order	3.84 (0.19)	5.64 (0.21)	4.95 (0.18)
Elaborative Interrogation	7.45 (0.18)	7.52 (0.16)	8.06 (0.14)
Self-explanation	6.61 (0.19)	6.90 (0.17)	7.41 (0.14)
Distributed Restudy Schedule	8.79 (0.12)	8.45 (0.14)	8.63 (0.12)
Retrieval Practice	8.79 (0.12)	8.59 (0.13)	8.15 (0.13)
Overt Retrieval Practice	8.97 (0.11)	-	8.16 (0.13)
Covert Retrieval Practice	8.03 (0.15)	-	7.52 (0.15)

Note: Ratings were made on a scale from 1 (*Very Ineffective*) to 10 (*Very Effective*). Standard errors of the mean are in parentheses. Only students and teachers made ratings for overt and covert retrieval practice.

3.1.1. Retrieval Practice vs. Highlighting

Significant main effects of strategy, $F(1,482) = 118.60$, $p < 0.001$, $\eta_p^2 = 0.20$, and group, $F(2,482) = 14.08$, $p < 0.001$, $\eta_p^2 = 0.06$, were qualified by a significant interaction between strategy and group, $F(2,482) = 11.49$, $p < 0.001$, $\eta_p^2 = 0.05$. Specifically, although all groups provided higher ratings of effectiveness for retrieval practice relative to highlighting, the difference was largest for students ($d = 0.86$), followed by teachers ($d = 0.64$) and parents ($d = 0.30$).

3.1.2. Retrieval Practice vs. Rereading

All groups provided higher ratings for retrieval practice ($M = 8.51$, $SE = 0.08$) compared to rereading ($M = 7.22$, $SE = 0.10$), $F(1,482) = 133.89$, $p < 0.001$, $\eta_p^2 = 0.22$. There was also a significant main effect of group, $F(2,482) = 13.90$, $p < 0.001$, $\eta_p^2 = 0.06$, indicating that students ($M = 8.11$, $SE = 0.12$) and parents ($M = 8.11$, $SE = 0.13$) gave higher ratings compared to teachers ($M = 7.37$, $SE = 0.11$), ps < 0.001. Students' and parents' ratings did not differ, $p = 1.00$. The interaction between strategy and group was not significant, $F(2,482) = 7.80$, $p = 0.07$, $\eta_p^2 = 0.01$.

3.1.3. Retrieval Practice vs. Keyword Mnemonic

Significant main effects of strategy, $F(1,482) = 80.05$, $p < 0.001$, $\eta_p^2 = 0.14$, and group, $F(2,482) = 14.49$, $p < 0.001$, $\eta_p^2 = 0.04$, were qualified by a significant interaction between strategy and group, $F(2,482) = 3.32$, $p = 0.04$, $\eta_p^2 = 0.01$. Specifically, although all groups rated retrieval practice as more effective compared to the keyword mnemonic, the difference was largest for parents ($d = 0.65$), followed by teachers ($d = 0.44$) and students ($d = 0.39$).

3.1.4. Retrieval Practice vs. Imagery

A significant main effect of strategy, $F(1,481) = 38.32$, $p < 0.001$, $\eta_p^2 = 0.07$, and a non-significant main effect of group, $F(2,481) = 2.47$, $p = 0.09$, $\eta_p^2 = 0.01$, were qualified by a significant interaction between strategy and group, $F(2,481) = 5.13$, $p = 0.006$, $\eta_p^2 = 0.02$.

Students and parents both provided higher ratings for retrieval practice compared to imagery, *p*s < 0.004. Teachers' ratings did not differ for these strategies, $p = 0.08$[10].

3.1.5. Retrieval Practice vs. Summarization

A significant main effect of strategy, $F(1,482) = 15.51$, $p < 0.001$, $\eta_p^2 = 0.03$, and a non-significant main effect of group, $F(2,482) = 1.72$, $p = 0.18$, $\eta_p^2 = 0.01$, were qualified by a significant interaction between strategy and group, $F(2,482) = 7.55$, $p < 0.001$, $\eta_p^2 = 0.03$. Both students and parents provided higher ratings for retrieval practice relative to summarization, *p*s < 0.001. Teachers' ratings did not differ between these strategies, $p = 0.40$.

3.1.6. Retrieval Practice vs. Massed Restudy Schedule

All groups gave higher ratings of effectiveness to retrieval practice ($M = 8.51$, $SE = 0.08$) relative to using a massed restudy schedule ($M = 4.61$, $SE = 0.11$), $F(1,482) = 892.36$, $p < 0.001$, $\eta_p^2 = 0.65$. A significant main effect of group, $F(2,482) = 12.83$, $p < 0.001$, $\eta_p^2 = 0.05$, indicated that both students ($M = 6.77$, $SE = 0.12$) and parents ($M = 6.80$, $SE = 0.12$) gave overall higher ratings compared to teachers ($M = 6.11$, $SE = 0.10$), *p*s < 0.001. Students' and parents' overall ratings did not differ, $p = 1.00$. The interaction was not significant, $F(2,482) = 1.51$, $p = 0.22$, $\eta_p^2 = 0.01$.

3.1.7. Retrieval Practice vs. Blocked Content Order

Participants' ratings of effectiveness were higher for retrieval practice ($M = 8.51$, $SE = 0.08$) compared to blocking content order ($M = 7.48$, $SE = 0.10$), $F(1,482) = 80.32$, $p < 0.001$, $\eta_p^2 = 0.14$. There was also a significant main effect of group, $F(2,482) = 4.16$, $p = 0.02$, $\eta_p^2 = 0.02$, which indicated that parents ($M = 8.15$, $SE = 0.12$) gave overall higher ratings compared to teachers ($M = 7.75$, $SE = 0.10$), $p = 0.03$. Students' ($M = 8.09$, $SE = 0.11$) ratings did not differ from parents', $p = 1.00$, or teachers', $p = 0.08$. The interaction was not significant, $F(2,482) = 2.63$, $p = 0.07$, $\eta_p^2 = 0.01$.

3.1.8. Retrieval Practice vs. Interleaved Content Order

Significant main effects of strategy, $F(1,482) = 720.73$, $p < 0.001$, $\eta_p^2 = 0.60$, and group, $F(2,482) = 49.30$, $p < 0.001$, $\eta_p^2 = 0.05$, were qualified by a significant interaction between strategy and group, $F(2,482) = 20.00$, $p < 0.001$, $\eta_p^2 = 0.08$. This interaction revealed that the difference in ratings between strategies was greatest for students ($d = 1.56$), followed by teachers ($d = 1.20$) and parents ($d = 1.14$).

3.1.9. Retrieval Practice vs. Elaborative Interrogation

A significant main effect of strategy, $F(1,481) = 56.41$, $p < 0.001$, $\eta_p^2 = 0.11$, and a non-significant main effect of group, $F(2,481) = 0.10$, $p = 0.91$, $\eta_p^2 = 0.00$, were qualified by a significant interaction between strategy and group, $F(2,481) = 12.62$, $p < 0.001$, $\eta_p^2 = 0.05$. Both students and parents gave higher ratings to retrieval practice relative to elaborative interrogation, *p*s < 0.001. By contrast, teachers' ratings did not differ, $p = 0.59$[11].

3.1.10. Retrieval Practice vs. Self-Explanation

A significant main effect of strategy, $F(1,482) = 185.18$, $p < 0.001$, $\eta_p^2 = 0.29$, and a non-significant main effect of group, $F(2,482) = 0.15$, $p = 0.86$, $\eta_p^2 = 0.001$, were qualified by a significant interaction between strategy and group, $F(2,482) = 15.22$, $p < 0.001$, $\eta_p^2 = 0.06$. The interaction demonstrated that although all groups rated retrieval practice as more effective compared to self-explanation, this difference was largest for students ($d = 0.98$), followed by parents ($d = 0.85$) and teachers ($d = 0.38$).

3.1.11. Retrieval Practice vs. Distributed Restudy Schedule

A significant main effect of group, $F(2,482) = 3.95$, $p = 0.02$, $\eta_p^2 = 0.02$, and a non-significant main effect of strategy, $F(1,482) = 1.69$, $p = 0.19$, $\eta_p^2 = 0.003$, were qualified by a significant interaction between strategy and group, $F(2,482) = 5.03$, $p = 0.007$, $\eta_p^2 = 0.02$. The interaction revealed that students' and parents' ratings of effectiveness for retrieval practice and using a distributed restudy schedule did not differ, ps > 0.38, whereas teachers rated using a distributed restudy schedule as more effective compared to using retrieval practice, $p < 0.001$.

3.1.12. Overt Retrieval Practice. vs. Covert Retrieval Practice

Overall, participants rated overt retrieval practice ($M = 8.56$, $SE = 0.09$) as more effective compared to covert retrieval practice ($M = 7.77$, $SE = 0.11$), $F(1, 339) = 72.80$, $p < 0.001$, $\eta_p^2 = 0.18$. There was also a significant main effect of group, $F(1,339) = 15.06$, $p < 0.001$, $\eta_p^2 = 0.04$, indicating that students ($M = 8.50$, $SE = 0.13$) rated these strategies as more effective than did teachers ($M = 7.84$, $SE = 0.11$). The interaction between strategy and group was not significant, $F(1,339) = 2.49$, $p = 0.12$, $\eta_p^2 = 0.007$.

3.1.13. Blocked Content Order vs. Interleaved Content Order

Significant main effects of strategy, $F(1,482) = 346.14$, $p < 0.001$, $\eta_p^2 = 0.42$, and group, $F(2,482) = 14.94$, $p < 0.001$, $\eta_p^2 = 0.06$, were qualified by a significant interaction between strategy and group, $F(2,482) = 9.26$, $p < 0.001$, $\eta_p^2 = 0.04$. Although all groups gave higher ratings of effectiveness for blocking content order relative to interleaving, this difference was largest for students ($d = 1.23$), followed by teachers ($d = 0.93$) and parents ($d = 0.82$).

3.1.14. Distributed Restudy Schedule vs. Massed Restudy Schedule

Significant main effects of strategy, $F(1,482) = 876.57$, $p < 0.001$, $\eta_p^2 = 0.65$, and group, $F(2,482) = 5.21$, $p = 0.006$, $\eta_p^2 = 0.02$, were qualified by a significant interaction between strategy and group, $F(2,482) = 26.25$, $p = 0.003$, $\eta_p^2 = 0.02$. Although all participants reported that using a distributed restudy schedule was more effective compared to using a massed restudy schedule, this difference was largest for teachers ($d = 1.51$), followed by students ($d = 1.45$) and parents ($d = 1.26$).

3.2. Ratings of Familiarity

Across groups, familiarity ratings were generally highest for retrieval practice and highlighting (see Table 5). By contrast, participants reported being less familiar with strategies such as interleaving and self-explanation, both of which can be effective for learning in some contexts. Lower familiarity with these strategies may partially explain why participants tended to give lower estimates of effectiveness for them compared to the other strategies.

Table 5. Ratings of Familiarity.

Learning Strategy	Students (n = 151)	Parents (n = 143)	Teachers (n = 191)
Highlighting	8.81 (0.14)	9.03 (0.14)	8.88 (0.12)
Rereading	8.36 (0.15)	8.57 (0.14)	8.38 (0.14)
Keyword Mnemonic	8.34 (0.18)	7.20 (0.23)	7.39 (0.18)
Imagery	7.64 (0.18)	8.01 (0.18)	7.90 (0.14)
Summarization	7.51 (0.20)	7.82 (0.20)	8.76 (0.11)
Massed Restudy Schedule	7.48 (0.20)	7.15 (0.21)	7.72 (0.16)
Blocked Content Order	7.66 (0.19)	8.22 (0.19)	8.24 (0.14)

Table 5. *Cont.*

Learning Strategy	Students (n = 151)	Parents (n = 143)	Teachers (n = 191)
Interleaved Content Order	3.07 (0.20)	5.50 (0.25)	4.31 (0.21)
Elaborative Interrogation	6.74 (0.20)	7.11 (0.21)	7.84 (0.17)
Self-explanation	5.97 (0.22)	6.48 (0.21)	7.33 (0.18)
Distributed Restudy Schedule	8.52 (0.14)	8.61 (0.15)	8.75 (0.13)
Retrieval Practice	9.15 (0.11)	8.77 (0.16)	8.91 (0.11)
Overt Retrieval Practice	9.05 (0.13)	-	8.60 (0.14)
Covert Retrieval Practice	8.88 (0.14)	-	8.56 (0.14)

Note: Ratings were made on a scale from 1 (*Very Unfamiliar*) to 10 (*Very Familiar*). Standard errors of the mean are in parentheses. Only students and teachers made ratings for overt and covert retrieval practice.

3.2.1. Retrieval Practice vs. Highlighting

Non-significant main effects of strategy, $F(1,482) = 0.14$, $p = 0.71$, $\eta_p^2 = 0.00$, and group, $F(2,482) = 0.19$, $p = 0.82$, $\eta_p^2 = 0.001$, were qualified by a significant interaction between strategy and group, $F(2,482) = 3.95$, $p = 0.02$, $\eta_p^2 = 0.02$. Students were more familiar with retrieval practice compared to highlighting, $p = 0.02$, whereas parents' and teachers' ratings of familiarity for these strategies did not differ, ps > 0.09.

3.2.2. Retrieval Practice vs. Rereading

A significant main effect of strategy, $F(1,482) = 30.98$, $p < 0.001$, $\eta_p^2 = 0.06$, and a non-significant main effect of group, $F(2,482) = 0.27$, $p = 0.77$, $\eta_p^2 = 0.001$, were qualified by a significant interaction between strategy and group, $F(2,482) = 3.30$, $p = 0.04$, $\eta_p^2 = 0.01$. Both students and teachers reported being more familiar with retrieval practice relative to rereading, ps < 0.001. By contrast, parents' ratings of familiarity did not differ, $p = 0.24$.

3.2.3. Retrieval Practice vs. Keyword Mnemonic

Significant main effects of strategy, $F(1,482) = 122.41$, $p < 0.001$, $\eta_p^2 = 0.20$, and group, $F(2,482) = 8.81$, $p < 0.001$, $\eta_p^2 = 0.04$, were qualified by a significant interaction between strategy and group, $F(2,482) = 4.25$, $p = 0.02$, $\eta_p^2 = 0.02$. Although all groups reported being more familiar with retrieval practice relative to the keyword mnemonic, this difference was largest for teachers ($d = 0.69$), followed by parents ($d = 0.63$) and students ($d = 0.43$).

3.2.4. Retrieval Practice vs. Imagery

A significant main effect of strategy, $F(1,481) = 106.46$, $p < 0.001$, $\eta_p^2 = 0.18$, and a non-significant main effect of group, $F(2,481) = 0.01$, $p = 1.00$, $\eta_p^2 = 0.00$, were qualified by a significant interaction between strategy and group, $F(2,481) = 4.06$, $p = 0.02$, $\eta_p^2 = 0.02$. Although all groups reported being more familiar with retrieval practice relative to imagery, the difference was largest for students ($d = 0.75$), followed by teachers ($d = 0.54$) and parents, ($d = 0.37$).

3.2.5. Retrieval Practice vs. Summarization

Significant main effects of strategy, $F(1,482) = 68.34$, $p < 0.001$, $\eta_p^2 = 0.12$, and group, $F(2,482) = 7.34$, $p < 0.001$, $\eta_p^2 = 0.03$, were qualified by a significant interaction between strategy and group, $F(2,482) = 16.31$, $p < 0.001$, $\eta_p^2 = 0.06$. Students and parents reported being more familiar with retrieval practice relative to summarization, ps < 0.001. Teachers' ratings of familiarity for these strategies did not differ, $p = 0.40$

3.2.6. Retrieval Practice vs. Massed Restudy Schedule

Overall, participants reported being more familiar with retrieval practice ($M = 8.94$, $SE = 0.07$) compared to using a massed restudy schedule ($M = 7.45$, $SE = 0.11$), $F(1,482) = 162.28$,

$p < 0.001$, $\eta_p^2 = 0.25$. The main effect of group was not significant, $F(2,482) = 2.56$, $p = 0.08$, $\eta_p^2 = 0.01$, nor was the interaction between strategy and group, $F(2,482) = 1.84$, $p = 0.16$, $\eta_p^2 = 0.01$.

3.2.7. Retrieval Practice vs. Blocked Content Order

A significant main effect of strategy, $F(1,482) = 75.83$, $p < 0.001$, $\eta_p^2 = 0.14$, and a non-significant main effect of group, $F(2,482) = 0.53$, $p = 0.59$, $\eta_p^2 = 0.002$, were qualified by a significant interaction between strategy and group, $F(2,482) = 7.76$, $p < 0.001$, $\eta_p^2 = 0.03$. Although all groups rated retrieval practice as more familiar compared to blocking content order, this difference was largest for students ($d = 0.73$), followed by teachers ($d = 0.37$) and parents ($d = 0.27$).

3.2.8. Retrieval Practice vs. Interleaved Content Order

Significant main effects of strategy, $F(1,482) = 965.76$, $p < 0.001$, $\eta_p^2 = 0.67$, and group, $F(2,482) = 14.82$, $p < 0.001$, $\eta_p^2 = 0.06$, were qualified by a significant interaction between strategy and group, $F(2,482) = 27.06$, $p < 0.001$, $\eta_p^2 = 0.10$. Although all groups reported being more familiar with retrieval practice relative to interleaving content order, this difference was largest for students ($d = 1.67$), followed by teachers ($d = 1.40$) and parents ($d = 1.09$).

3.2.9. Retrieval Practice vs. Elaborative Interrogation

Significant main effects of strategy, $F(1,481) = 194.57$, $p < 0.001$, $\eta_p^2 = 0.29$, and group, $F(2,481) = 4.38$, $p = 0.01$, $\eta_p^2 = 0.02$, were qualified by a significant interaction between strategy and group, $F(2,481) = 10.45$, $p < 0.001$, $\eta_p^2 = 0.04$. Although all groups reported being more familiar with retrieval practice relative to elaborative interrogation, this difference was largest for students ($d = 1.02$), followed by parents ($d = 0.70$) and teachers ($d = 0.51$).

3.2.10. Retrieval Practice vs. Self-Explanation

Significant main effects of strategy, $F(1,482) = 352.53$, $p < 0.001$, $\eta_p^2 = 0.42$, and group, $F(2,482) = 5.95$, $p = 0.003$, $\eta_p^2 = 0.02$, were qualified by a significant interaction between strategy and group, $F(2,482) = 14.43$, $p < 0.001$, $\eta_p^2 = 0.06$. Although all groups reported being more familiar with retrieval practice relative to self-explanation, this difference was largest for students ($d = 1.18$), followed by parents ($d = 0.91$) and teachers ($d = 0.72$).

3.2.11. Retrieval Practice vs. Distributed Restudy Schedule

A significant main effect of strategy was also significant, $F(1,482) = 13.90$, $p < 0.001$, $\eta_p^2 = 0.03$, and a non-significant main effect of group, $F(2,482) = 0.48$, $p = 0.62$, $\eta_p^2 = 0.002$, was qualified by a significant interaction between strategy and group, $F(2,482) = 3.36$, $p = 0.04$, $\eta_p^2 = 0.01$. The interaction revealed that students reported being more familiar with retrieval practice relative to using a distributed restudy schedule, $p < 0.001$, whereas parents' and teachers' ratings of familiarity for these strategies did not differ, ps > 0.24.

3.2.12. Overt Retrieval Practice. vs. Covert Retrieval Practice

Participants' ratings of familiarity for covert and overt retrieval practice did not differ, $F(1,339) = 1.36$, $p = 0.24$, $\eta_p^2 = 0.004$. A significant main effect of group, $F(1,339) = 4.75$, $p = 0.03$, $\eta_p^2 = 0.01$, indicated that students reported being more familiar with these strategies ($M = 8.96$, $SE = 0.13$) relative to teachers ($M = 8.58$, $SE = 0.12$). The interaction between strategy and group was not significant, $F(1,339) = 0.55$, $p = 0.46$, $\eta_p^2 = 0.002$.

3.2.13. Blocked Content Order vs. Interleaved Content Order

Significant main effects of strategy, $F(1,482) = 603.33$, $p < 0.001$, $\eta_p^2 = 0.56$, and group, $F(2,482) = 24.22$, $p < 0.001$, $\eta_p^2 = 0.09$, were qualified by a significant interaction between strategy and group, $F(2,482) = 11.81$, $p < 0.001$, $\eta_p^2 = 0.05$. Although all groups reported being more familiar with blocking content order compared to interleaving content order, this difference was largest for students ($d = 1.38$), followed by teachers ($d = 1.11$) and parents ($d = 0.92$).

3.2.14. Distributed Restudy Schedule vs. Massed Restudy Schedule

All groups reported being more familiar with using a distributed restudy schedule ($M = 8.63$, $SE = 0.08$) relative to using a massed restudy schedule ($M = 7.45$, $SE = 0.11$), $F(1,482) = 89.55$, $p < 0.001$, $\eta_p^2 = 0.16$. The main effect of group was not significant, $F(2,482) = 2.17$, $p = 0.12$, $\eta_p^2 = 0.01$, nor was the interaction, $F(2,482) = 1.26$, $p = 0.28$, $\eta_p^2 = 0.01$.

3.3. Ratings of Future Use or Recommendation

Across all groups, participants gave the highest ratings for using a distributed restudy schedule and retrieval practice (see Table 6). Further, participants in all groups gave the lowest ratings to interleaving content order and using a massed restudy schedule.

Table 6. Ratings of the Likelihood of Future Use or Recommendation.

Learning Strategy	Students ($n = 151$)	Parents ($n = 143$)	Teachers ($n = 191$)
Highlighting	7.38 (0.21)	8.03 (0.18)	6.80 (0.19)
Rereading	7.39 (0.20)	7.52 (0.18)	6.36 (0.19)
Keyword Mnemonic	7.87 (0.21)	6.87 (0.24)	6.81 (0.17)
Imagery	7.13 (0.20)	7.92 (0.16)	7.62 (0.14)
Summarization	6.89 (0.21)	7.69 (0.19)	8.01 (0.14)
Massed Restudy Schedule	5.48 (0.23)	4.43 (0.23)	3.57 (0.18)
Blocked Content Order	6.75 (0.21)	7.52 (0.20)	7.06 (0.17)
Interleaved Content Order	2.64 (0.18)	5.26 (0.24)	3.96 (0.20)
Elaborative Interrogation	6.29 (0.21)	7.37 (0.20)	7.78 (0.16)
Self-explanation	5.55 (0.22)	6.46 (0.22)	7.11 (0.17)
Distributed Restudy Schedule	7.93 (0.16)	8.38 (0.14)	8.54 (0.13)
Retrieval Practice	8.62 (0.13)	8.50 (0.15)	8.02 (0.15)
Overt Retrieval Practice	8.40 (0.16)	-	7.94 (0.15)
Covert Retrieval Practice	8.01 (0.17)	-	7.19 (0.17)

Note: Ratings were made on a scale from 1 (*Very Unlikely*) to 10 (*Very Likely*). Standard errors of the mean are in parentheses. Only students and teachers made ratings for overt and covert retrieval practice.

3.3.1. Retrieval Practice vs. Highlighting

Significant main effects of strategy, $F(1,482) = 56.46$, $p < 0.001$, $\eta_p^2 = 0.11$, and group, $F(2,482) = 3.62$, $p = 0.03$, $\eta_p^2 = 0.02$, were qualified by a significant interaction between strategy and group, $F(2,482) = 3.62$, $p = 0.03$, $\eta_p^2 = 0.02$. Both students, $p < 0.001$, and teachers, $p < 0.001$, gave higher ratings for retrieval practice compared to highlighting. However, parents' recommendation ratings did not significantly differ between retrieval practice and highlighting, $p = 0.73$.

3.3.2. Retrieval Practice vs. Rereading

All groups provided higher future use and recommendation ratings for retrieval practice ($M = 8.34$, $SE = 0.08$) as compared to rereading ($M = 7.02$, $SE = 0.11$), $F(1,482) = 99.10$, $p < 0.001$, $\eta_p^2 = 0.17$. There was a significant effect of group, $F(2,482) = 15.07$, $p < 0.001$, $\eta_p^2 = 0.06$[12], such that teachers ($M = 7.19$, $SE = 0.12$) provided lower ratings than did students ($M = 8.00$, $SE = 0.13$), $p < 0.001$, and parents ($M = 8.01$, $SE = 0.13$), $p < 0.001$.

Students' and parents' ratings did not differ, $p = 1.00$. The interaction between strategy and group was not significant, $F(2,482) = 2.50$, $p = 0.09$, $\eta_p^2 = 0.01$.

3.3.3. Retrieval Practice vs. Keyword Mnemonic

Significant main effects of strategy, $F(1,482) = 84.70$, $p < 0.001$, $\eta_p^2 = 0.15$, and group, $F(2,482) = 9.96$, $p < 0.001$, $\eta_p^2 = 0.04$, were qualified by a significant interaction between strategy and group, $F(2,482) = 3.58$, $p = 0.029$, $\eta_p^2 = 0.02$. Although all groups gave higher ratings to retrieval practice relative to the keyword mnemonic, this difference was largest for parents ($d = 0.65$), followed by teachers ($d = 0.52$) and students ($d = 0.35$).

3.3.4. Retrieval Practice vs. Imagery

A significant main effect of strategy, $F(1,482) = 50.19$, $p < 0.001$, $\eta_p^2 = 0.09$, and a non-significant main effect of group, $F(2,482) = 2.93$, $p = 0.06$, $\eta_p^2 = 0.01$, were qualified by a significant interaction between strategy and group, $F(2,482) = 8.28$, $p < 0.001$, $\eta_p^2 = 0.03$. Students indicated higher future use ratings for retrieval practice as compared to imagery, $p < 0.001$. By contrast, both parents, $p = 0.10$, and teachers, $p = 0.40$, provided similar recommendation ratings for retrieval practice and imagery.

3.3.5. Retrieval Practice vs. Summarization

A significant main effect of strategy, $F(1,482) = 49.01$, $p < 0.001$, $\eta_p^2 = 0.09$, and a non-significant main effect of group, $F(2,482) = 2.01$, $p = 0.14$, $\eta_p^2 = 0.01$, were qualified by a significant interaction between strategy and group, $F(2,482) = 17.79$, $p < 0.001$, $\eta_p^2 = 0.07$. The interaction revealed that students and parents provided higher ratings for retrieval practice as compared to summarization, ps < 0.005. Teachers did not rate retrieval practice significantly differently than summarization, $p = 1.00$.

3.3.6. Retrieval Practice vs. Massed Restudy Schedule

Significant main effects of strategy, $F(1,482) = 699.06$, $p < 0.001$, $\eta_p^2 = 0.59$, and group, $F(2,482) = 26.02$, $p < 0.001$, $\eta_p^2 = 0.10$, were qualified by a significant interaction between strategy and group, $F(2,482) = 37.17$, $p < 0.001$, $\eta_p^2 = 0.03$. Although all groups gave higher ratings to retrieval practice compared to using a massed restudy schedule, the interaction revealed that this difference was largest for teachers ($d = 1.41$), followed by parents ($d = 1.33$) and students ($d = 1.13$).

3.3.7. Retrieval Practice vs. Blocked Content Order

Significant main effects of strategy, $F(1,482) = 90.09$, $p < 0.001$, $\eta_p^2 = 0.16$, and group, $F(2,482) = 3.64$, $p = 0.03$, $\eta_p^2 = 0.02$, were qualified by a significant interaction between strategy and group, $F(2,482) = 5.05$, $p = 0.007$, $\eta_p^2 = 0.02$. The interaction revealed that although all groups gave higher ratings to retrieval practice compared to blocking content order, this difference was largest for students ($d = 0.80$), followed by parents ($d = 0.44$) and teachers ($d = 0.43$).

3.3.8. Retrieval Practice vs. Interleaved Content Order

Significant main effects of strategy, $F(1,482) = 893.87$, $p < 0.001$, $\eta_p^2 = 0.65$, and group, $F(2,482) = 25.30$, $p < 0.001$, $\eta_p^2 = 0.10$, were qualified by a significant interaction between strategy and group, $F(2,482) = 28.32$, $p < 0.001$, $\eta_p^2 = 0.11$. Although all groups gave higher ratings to retrieval practice compared to interleaving content order, this difference was largest for students ($d = 1.69$), followed by teachers ($d = 1.30$) and parents ($d = 1.12$).

3.3.9. Retrieval Practice vs. Elaborative Interrogation

Significant main effects of strategy, $F(1,482) = 84.32$, $p < 0.001$, $\eta_p^2 = 0.15$, and group, $F(2,482) = 4.84$, $p = 0.008$, $\eta_p^2 = 0.02$, were qualified by a significant interaction between strategy and group, $F(2,482) = 21.31$, $p < 0.001$, $\eta_p^2 = 0.08$. Students and parents provided higher ratings for retrieval practice as compared to elaborative interrogation, ps < 0.001. Teachers did not rate retrieval practice significantly differently than elaborative interrogation, $p = 1.00$.

3.3.10. Retrieval Practice vs. Self-Explanation

Significant main effects of strategy, $F(1,482) = 208.81$, $p < 0.001$, $\eta_p^2 = 0.30$, and group, $F(2,482) = 4.10$, $p = 0.02$, $\eta_p^2 = 0.02$, were qualified by a significant interaction between strategy and group, $F(2,482) = 21.64$, $p < 0.001$, $\eta_p^2 = 0.08$. Although all groups provided higher ratings for retrieval practice compared to self-explanation, the interaction revealed that this difference was largest for students ($d = 1.15$), followed by parents ($d = 0.84$) and teachers ($d = 0.40$).

3.3.11. Retrieval Practice vs. Distributed Restudy Schedule

Non-significant main effects of strategy, $F(1,482) = 0.82$, $p = 0.37$, $\eta_p^2 = 0.002$, and group, $F(2,482) = 0.62$, $p = 0.54$, $\eta_p^2 = 0.003$, were qualified by a significant interaction between strategy and group, $F(2,482) = 13.73$, $p < 0.001$, $\eta_p^2 = 0.05$. The interaction revealed that students provided higher future use ratings for retrieval practice as compared to a distributed restudy schedule, $p < 0.001$. By contrast, teachers provided higher recommendation ratings for a distributed restudy schedule as compared to retrieval practice, $p = 0.01$. Parents' ratings did not significantly differ between retrieval practice and a distributed restudy schedule, $p = 1.00$.

3.3.12. Overt Retrieval Practice. vs. Covert Retrieval Practice

Participants provided higher future use and recommendation ratings for overt retrieval practice ($M = 8.17$, $SE = 0.11$) as compared to covert retrieval practice ($M = 7.56$, $SE = 0.12$), $F(1,339) = 28.38$ $p < 0.001$, $\eta_p^2 = 0.08$. There was also a significant effect of group, $F(1,339) = 9.76$, $p = 0.002$, $\eta_p^2 = 0.03$, such that students ($M = 8.21$, $SE = 0.15$) provided higher ratings than did teachers ($M = 7.57$, $SE = 0.14$). The interaction between strategy and group was not significant, $F(1,339) = 2.79$, $p = 0.10$, $\eta_p^2 = 0.01$.

3.3.13. Blocked Content Order vs. Interleaved Content Order

Significant main effects of strategy, $F(1,482) = 382.66$, $p < 0.001$, $\eta_p^2 = 0.44$, and group, $F(2,482) = 32.98$, $p < 0.001$, $\eta_p^2 = 0.12$, were qualified by a significant interaction between strategy and group, $F(2,482) = 10.10$, $p < 0.001$, $\eta_p^2 = 0.04$. Although all groups gave higher ratings for a blocked content order as compared to an interleaved content order, this difference was largest for students ($d = 1.31$), followed by teachers ($d = 1.05$) and parents ($d = 0.79$).

3.3.14. Distributed Restudy Schedule vs. Massed Restudy Schedule

Significant main effects of strategy, $F(1,482) = 598.00$, $p < 0.001$, $\eta_p^2 = 0.55$, and group, $F(2,482) = 8.08$, $p < 0.001$, $\eta_p^2 = 0.03$, were qualified by a significant interaction between strategy and group, $F(2,482) = 23.28$, $p < 0.001$, $\eta_p^2 = 0.09$. Although all groups provided higher ratings for using a distributed restudy schedule relative to using a massed restudy schedule, the interaction revealed that this difference was largest for teachers ($d = 1.52$), followed by parents ($d = 1.32$) and students ($d = 0.91$).

4. Discussion

We evaluated students', teachers', and parents' knowledge and perceptions about several common learning strategies by providing participants with concrete examples of how each strategy could be implemented and asking them to rate each on effectiveness, familiarity, and their likelihood of using or recommending it.

Prior research suggested that students typically underestimated the impact of highly effective strategies such as using retrieval practice and distributing study sessions over time (Blasiman et al., 2017; McCabe, 2011). By contrast, our outcomes demonstrated that students (and parents and teachers) hold accurate knowledge about these strategies. These strategies received the highest effectiveness ratings, and participants generally rated retrieval practice as more effective compared to all other strategies. In addition, when directly comparing retrieval practice and using a distributed restudy schedule, there were minimal differences in participants' ratings of effectiveness. This difference between the outcomes of the present study and those of prior research may have arisen due to the differences in methodology. For instance, in the present research, participants were allowed to rate each strategy independently, rather than pitting it against another strategy that they may have also believed was effective. When rating each strategy in isolation, the students provided higher ratings for effective strategies (such as self-testing) compared to less effective strategies (such as rereading), whereas in research where these strategies were pitted against each other, the students reported rereading to be more effective than self-testing (McCabe, 2011). Using a paradigm where participants compared two different strategies may have biased their ratings because of the evaluative comparison of those two strategies. Relatedly, Blasiman et al. (2017) had students rate individual learning strategies in isolation and found that the students provided lower ratings of effectiveness for practice testing compared to other less effective strategies (e.g., reading notes, copying notes, highlighting notes), which contrasts with the present outcomes. However, the study by Blasiman et al. (2017) did not provide concrete examples for strategy implementation. Given that students sometimes view retrieval practice as a monitoring tool rather than a learning strategy, this may explain why the students in this study provided lower ratings of effectiveness for it as a learning strategy (Kornell & Bjork, 2007). More optimistically, differences between the present study and past research may also have been driven by greater dissemination of evidenced-based learning practices through popular media since the publication of those studies (2011 and 2017). For instance, books like *Make it Stick: The Science of Successful Learning* (Brown et al., 2014) and *Powerful Teaching: Unleash the Science of Learning* (Agarwal & Bain, 2019) have made research outcomes on education more accessible. This may have had the desired goal of leading people to become more informed about effective learning strategies.

Although students appeared to have accurate knowledge about effective strategies, they also appeared to have some inaccurate knowledge, overestimating the effectiveness of some strategies that are less effective for learning and comprehension (e.g., highlighting rereading). Even though participants' ratings for these strategies was statistically lower compared to retrieval practice, numerically, they still received relatively high ratings (all above 6.5 on a 10-point scale). This outcome could have negative implications for students' learning. Strategies such as rereading and highlighting are less effortful compared to the more effective strategies like retrieval practice. As such, if students think these strategies are somewhat effective, they may choose to use them more often than retrieval practice. This may be especially likely because students' perceived time costs introduce a significant barrier for their decisions to use more effective learning strategies (Rea et al., 2022). In addition, the students also provided very low effectiveness ratings for interleaving content order during study (significantly lower than both retrieval practice and blocking content

order), which in some contexts can be an effective strategy (Hartwig et al., 2022). Familiarity ratings may give insight into this outcome. Specifically, students (and teachers and parents) appear to be relatively unfamiliar with interleaving as a learning strategy. Thus, they may think it is an ineffective strategy simply because they have never learned about it.

In addition to developing a better understanding of students' knowledge about learning strategies, key contributions of the present research are the outcomes relating to teachers and parents. Although teachers and parents play a key role in students' development of knowledge about learning strategies, surprisingly little research has evaluated the accuracy of their knowledge about learning strategies (Lieder et al., 2023). The outcomes from these groups were largely consistent with those reported with students. The teachers and parents provided high ratings of effectiveness for retrieval practice and using a distributed restudy schedule, yet still provided relatively high ratings for some of the less effective strategies (e.g., highlighting and rereading). This could be problematic if teachers and parents recommend less effective strategies to their students/children. Both groups reported being most likely to recommend retrieval practice and distributed restudy schedules, which is promising. Even so, they also reported being almost as likely to recommend less effective strategies such as highlighting (parents) and summarization (teachers). A key goal for education researchers will be to find ways to broadly disseminate their research to provide access to teachers and parents so that they can update their knowledge accordingly. In turn, teachers and parents will be able to make evidence-based recommendations to their students and children.

The outcomes of the present research have important applied implications. First, the methodological choices we made allowed us to evaluate participants' knowledge of strategies in isolation, which may more clearly mirror how students consider them in real learning contexts. For instance, when making study decisions when studying for classes, students are unlikely to have two strategies pitted against each other from which they choose which to use. Instead, students likely rely on their knowledge of a particular strategy to determine whether they will use it. This decision can be influenced by several factors (e.g., how effective they think it is, whether they have used it before, if it was recommended to them, if they have enough time to use it). Thus, the way we measured students' knowledge may be more likely to generalize to applied contexts compared to methodologies that have pitted strategies against each other.

A second applied implication relates to participants' familiarity with the strategies. The present data suggest that, for each group of participants, there is variability in people's familiarity with these strategies. Thus, simply telling students (or teachers and parents) that a strategy (e.g., self-explanation) is effective may not be enough if they are unfamiliar with the strategy and do not understand how to implement it appropriately. Instead, the present outcomes suggest that when educating people about learning strategies (e.g., teachers telling students which strategies are effective; professional development workshop leaders educating teachers on effective strategies), they need to be instructed how to effectively implement them.

An important consideration for interpreting the outcomes of the present research is to note that research evaluating the effectiveness of these strategies is ongoing. Whereas the effects of some of these strategies appear to be quite robust (e.g., retrieval practice is highly effective in many contexts), for other strategies, the effectiveness has been debated (Dunlosky et al., 2013). For example, in many instances, interleaving content during learning is beneficial, but, for some types of materials, blocked content during learning is ideal (Carvalho & Goldstone, 2015). Additionally, the effectiveness of the strategy may depend on the learner's goals. For instance, interleaving may be more effective when the goal is to identify differences between items, whereas blocking may be more effective

when the goal is to identify similarities amongst items (Yan et al., 2017). Moreover, the effectiveness of some strategies (e.g., self-explanation; highlighting) can depend on the way those strategies are used and if they are used in combination with other strategies (Miyatsu et al., 2018; Ponce et al., 2022). Our goal was not to set definitive guidelines for the effectiveness of each strategy but rather to evaluate students', teachers', and parents' knowledge and perceptions of them. Thus, we encourage readers to think flexibly about the effectiveness of these strategies as empirical work continues.

Finally, it is also worth noting that each concrete example we provided participants with for each strategy was only one example of how a strategy could be implemented. Many of the strategies we investigated could be implanted in several different ways, and the ways that those strategies are implemented could moderate their effectiveness. For instance, our scenario on interleaving focused on interleaving content from two different learning domains. Implementing this strategy in this way can be less effective for learning compared to interleaving content order within a given domain (e.g., for learning differences between different related concepts; Hausman & Kornell, 2014). As another example, although Dunlosky et al. (2013) gave the keyword mnemonic a low utility rating, it can be effective when used for certain types of materials (e.g., foreign language vocabulary; Pressley et al., 1982). Given that our concrete example for the keyword mnemonic involved learning foreign language vocabulary (see Table 2), this could explain why participants in our study tended to rate it relatively high in effectiveness. Had we chosen a different example (e.g., learning a text passage), participants may have lowered their estimates of effectiveness. Accordingly, interpretation of participants' ratings in the present study should be contextualized to the specific example of implementation provided rather than to other ways that strategy could be implemented. We encourage additional research to continue exploring students', parents', and teachers' metacognitive knowledge related to study strategies, specifically on how the way that a strategy is implemented might impact their beliefs about the effectiveness of that strategy.

As another limitation, some of the specific wording used in our examples may have been less clear than we originally intended. This may have led to idiosyncratic differences in participants' interpretations of our example scenarios. For instance, for the rereading scenario, we provided two possible ways to implement the strategy. One mentioned revisiting a map to study to location of the United States, whereas the other mentioned rereading a textbook to learn about each state. Given that there are multiple ways to consider the word revisiting (e.g., rereading, self-testing, distributing study), it is possible that there was ambiguity in how participants interpreted this scenario, and some participants may have considered multiple strategies when making their ratings. Thus, the wording of some scenarios, and the specific example topics chosen, pose limitations that we encourage readers to consider when contextualizing our results. To address this limitation, future research should (a) carefully evaluate the wording chosen to ensure that there is no ambiguity and (b) counterbalance the example topics chosen across strategies to ensure that participants' ratings are representative of the strategy and not specific to the to-be-learned material in the example.

5. Conclusions

In sum, the present research revealed several novel outcomes that contribute to the literature regarding students', teachers', and parents' knowledge about learning strategies. The accuracy of people's knowledge appears to be mixed with some evidence of accurate knowledge (e.g., for strategies high in effectiveness) and some evidence of inaccurate knowledge (e.g., for strategies low in effectiveness). Studies that investigate people's metacognitive knowledge are critical because students', teachers', and parents' knowledge

can influence the decisions that students make about how to study, which can impact their learning.

Author Contributions: Conceptualization, A.E.W., A.L.B. and S.K.T.; methodology, A.E.W. and S.K.T.; materials, A.E.W. and S.K.T.; formal analysis, A.E.W. and A.L.B.; writing—original draft preparation, A.E.W.; writing—review and editing, A.E.W., A.L.B. and S.K.T.; project administration, A.E.W. and A.L.B.; funding acquisition, S.K.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition, collaborative grant number 220020483.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Texas Christian University (protocol code 1709-014-1709 and 29 June 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All materials, raw data, and group comparison analyses have been uploaded to the Open Science Framework and can be accessed at https://osf.io/ghtyn/. The study's design and analyses were not pre-registered.

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

1 One notable exception is a recently published article by Wissman and Leontyev (2024) which measured teachers' knowledge about six learning strategies (only three of which were also explored in the present study) in isolation and by giving concrete examples of each. Consistent with the research discussed, they found that teachers' knowledge about the strategies they explored was low.

2 To be eligible to participate in the survey mTurk workers (parents and teachers) were required to have a HIT approval rate of 95–100% and have fewer than 500 HITs approved.

3 We conducted an unrelated experiment with children aged 10–17. Parents who brought their children in for that experiment were recruited to complete the present survey.

4 We chose not to include these participants in our teaching sample because they did not receive the teacher version of the survey, and as such, we did not have information about their teaching experience. As well, the questions they answered in the parent version of the survey were framed around their child, and it is unclear whether their responses might differ if the questions were framed around their students.

5 For these participants, we e-mailed recruitment flyers to administrative assistants in STEM departments at colleges and universities in Texas as well as local middle schools and high schools.

6 The numbers sum to greater than 191 because some teachers reported teaching multiple topics

7 The teacher version of the survey found in the online supplemental materials (https://osf.io/ghtyn/) is the exact version that teachers who completed a physical copy received.

8 These questions are not directly related to the present research and were included to provide more data on issues related to students' and teachers' notetaking beliefs and practices. As such, we do not report their outcomes in the present manuscript. For interested readers, we provide the outcomes in Appendix A of the online supplemental materials (https://osf.io/ghtyn/).

9 For these comparisons, we could have alternatively chosen to include using a distributed restudy schedule as the comparison strategy, as there is ample evidence that it is also a highly effective strategy and was given a high utility rating by Dunlosky et al. (2013). Our choice to use retrieval practice was largely because it is more similar to the other strategies compared to using a distributed restudy schedule (or a massed restudy schedule). For instance, in any single study session, students could choose to use retrieval practice or any of the other strategies (e.g., imagery, rereading). As well, retrieval practice may be more likely to be adopted by students compared to using a distributed restudy schedule (for support for this, see Section 3.3.11) because it does not require them to plan out their study ahead of time. Most important, there would likely be minimal differences in our outcomes had we chosen to include using a distributed restudy schedule as the comparison instead of retrieval practice, as there were minimal differences in participants' ratings between these strategies (see Sections 3.1.11, 3.2.11 and 3.3.11) and both were given relatively high ratings on all measures.

10 One teacher did not provide any ratings for any of the imagery questions.

11 One teacher did not provide any ratings for any of the elaborative interrogation questions.

[12] The assumption of homogeneity of variance was violated for rereading (Levene's test: $F(2,482) = 5.43$, $p = 0.005$), thus, we computed a Welch's Robust Tests of Equality of Means. This analysis indicated that there was a significant effect of group even when accounting for unequal variances, $F(2, 316.73) = 11.66$, $p < 0.001$.

References

Agarwal, P. K., & Bain, P. M. (2019). *Powerful teaching: Unleash the science of learning*. Jossey Bass Incorporated.

Anthenien, A. M., DeLozier, S. J., Neighbors, C., & Rhodes, M. G. (2018). College student normative misperceptions of peer study habit use. *Social Psychology of Education*, *21*, 303–322. [CrossRef]

Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: Comparing intended versus actual study behaviour. *Memory*, *25*, 784–792. [CrossRef]

Brown, P. C., Roediger, H. L., III, & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. The Belknap Press of Harvard University Press.

Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, *22*, 281–288. [CrossRef]

Cheung, C. S.-S., & Pomerantz, E. M. (2012). Why does parents' involvement enhance children's achievement? The role of parent-oriented motivation. *Journal of Educational Psychology*, *104*, 820–832. [CrossRef]

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Sage Publications, Inc.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*, 4–58. [CrossRef] [PubMed]

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*, 717–741. [CrossRef]

Firth, J. W. (2021). Teachers' beliefs about memory: A vignette study of trainee and in-service teachers. *Studia Psychologica*, *63*, 204–220. [CrossRef]

Flavell, J. H. (1971). First discussant's comments: What is memory development the development of? *Human Development*, *14*(4), 272–278. [CrossRef]

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*, 906–911. [CrossRef]

Geller, J., Toftness, A. R., Armstrong, P. I., Carpenter, S. K., Manz, C. L., Coffman, C. R., & Lamm, M. H. (2018). Study strategies and beliefs about learning as a function of academic achievement and achievement goals. *Memory*, *26*, 683–690. [CrossRef] [PubMed]

Halamish, V. (2018). Pre-service and in-service teachers' metacognitive knowledge of learning strategies. *Frontiers in Psychology*, *9*, 1–5. [CrossRef]

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*, 126–134. [CrossRef]

Hartwig, M. K., Rohrer, D., & Dedrick, R. F. (2022). Scheduling math practice: Students' underappreciation of spacing and interleaving. *Journal of Experimental Psychology: Applied*, *28*, 100–113. [CrossRef]

Hausman, H., & Kornell, N. (2014). Mixing topics while studying does not enhance learning. *Journal of Applied Research in Memory and Cognition*, *3*, 153–160. [CrossRef]

Hoover-Dempsey, K. V., Bassler, O. C., & Burow, R. (1995). Parents' reported involvement in students' homework: Strategies and practices. *The Elementary School Journal*, *95*, 435–450. [CrossRef]

Hunter, A. S., & Lloyd, M. E. (2018). Faculty discuss study strategies, but not the best ones: A survey of suggested exam preparation techniques for difficult courses across disciplines. *Scholarship of Teaching and Learning in Psychology*, *4*, 105–114. [CrossRef]

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*, 219–224. [CrossRef]

Levin, J. R., Shriberg, L. K., Miller, G. E., McCormick, C. B., & Levin, B. B. (1980). The keyword method in the classroom: How to remember the states and their capitals. *The Elementary School Journal*, *80*, 185–191. [CrossRef]

Lieder, E. R., Nakazato, N., Ohtani, K., Ishii, R., Fukuzumi, N., Sakaki, M., Ishikawa, S., Suzuki, T., Murayama, K., & Tanaka, A. (2023). Children's study habits are predicted by their parents' learning strategy preferences. *Learning and Instruction*, *88*, 101809. [CrossRef]

Martinez-Pons, M. (2002). Parental influences on children's academic self-regulatory development. *Theory into Practice*, *41*, 126–131. [CrossRef]

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*, 462–476. [CrossRef]

Miyatsu, T., Nguyen, K., & McDaniel, M. A. (2018). Five popular study strategies: Their pitfalls and optimal implementations. *Perspectives on Psychological Science*, *13*, 390–407. [CrossRef]

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*, 257–271. [CrossRef] [PubMed]

Ponce, H. R., Mayer, R. E., & Méndez, E. E. (2022). Effects of learner-generated highlighting and instructor-provided highlighting on learning from text: A meta-analysis. *Educational Psychology Review*, *34*, 989–1024. [CrossRef]

Pressley, M., Levin, J. R., & Delaney, H. D. (1982). The mnemonic keyword method. *Review of Educational Research*, *52*, 61–91. [CrossRef]

Rea, S. D., Wang, L., Muenks, K., & Yan, V. X. (2022). Students can (mostly) recognize effective learning, so why do they not do it? *Journal of Intelligence*, *10*, 127. [CrossRef] [PubMed]

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*, 20–27. [CrossRef] [PubMed]

Tauber, S. K., & Dunlosky, J. (2016). A brief history of metamemory research and handbook overview. In J. Dunlosky, & S. K. Tauber (Eds.), *The Oxford handbook of Metamemory* (pp. 7–21). Oxford University Press. [CrossRef]

Tauber, S. K., Witherby, A. E., Dunlosky, J., Rawson, K. A., Putnam, A. L., & Roediger, H. L., III. (2018). Does covert retrieval benefit learning of key-term definitions? *Journal of Applied Research in Memory and Cognition*, *7*, 106–115. [CrossRef]

Wissman, K. T., & Leontyev, A. (2024). Awareness and implementation of evidence-based learning strategies among STEM faculty. *Journal of College Science Teaching*, *53*, 345–354. [CrossRef]

Witherby, A. E., & Tauber, S. K. (2019). The current status of students' note-taking: Why and how do students take notes? *Journal of Applied Research in Memory and Cognition*, *8*, 139–153. [CrossRef]

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*, 918–933. [CrossRef] [PubMed]

Yan, V. X., Soderstrom, N. C., Seneviratna, G. S., Bjork, E. L., & Bjork, R. A. (2017). How should exemplars be sequenced in inductive learning? Empirical evidence versus learners' opinions. *Journal of Experimental Psychology: Applied*, *23*, 403–416. [CrossRef] [PubMed]

Yan, V. X., Thai, K.-P., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: Do they vary with mindset? *Journal of Applied Research in Memory and Cognition*, *3*, 140–152. [CrossRef]

Zepeda, C. D., & Nokes-Malach, T. J. (2021). Metacognitive study strategies in a college course and their relation to exam performance. *Memory & Cognition*, *49*, 480–497. [CrossRef]

Zimmerman, B. J., & Pons, M. M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal*, *23*, 614–628. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Durability of Students' Learning Strategies Use and Beliefs Following a Classroom Intervention

Ezgi M. Yüksel [1,*], C. Shawn Green [1] and Haley A. Vlach [2]

[1] Department of Psychology, University of Wisconsin-Madison, Madison, WI 53706, USA; cshawn.green@wisc.edu
[2] Department of Educational Psychology, University of Wisconsin-Madison, Madison, WI 53706, USA; hvlach@wisc.edu
* Correspondence: yksel@wisc.edu; Tel.: +1-6084404502

Abstract

When students choose their own learning strategies, they often rely on ineffective methods, such as rereading and cramming, which have limited long-term benefits. To improve learning outcomes, previous interventions have utilized explicit instruction about effective strategies and direct experience with those strategies, though with mixed success. Yüksel et al. demonstrated that combining both approaches could foster initial improvements in students' understanding and use of effective learning strategies. In Study 1, we examined the long-term effects of this combined intervention by contacting participants six months later to assess the stability of outcomes. In Study 2, we extended the scope by surveying all students who had enrolled in the intervention section over the past five years. Participants were asked about their use and perceived effectiveness of various strategies. In both studies, quantitative measures were complemented with open-ended questions to gain deeper insights into study behaviors and obstacles to adopting effective strategies. While students retained an understanding of the effectiveness of various strategies and reported using ineffective strategies less frequently, the adoption of more effective strategies did not show a significant increase. However, compared to the business-as-usual group, the intervention group did not experience a decline in their use of effective strategies. These results suggest that while explicit instruction and experience can enhance knowledge, long-term behavior change remains difficult. Reported obstacles—such as time constraints, limited resources, procrastination, and prioritizing short-term gains—align with metacognitive theories of desirable difficulties and help explain why students still favor less effortful strategies, despite knowing more effective ones that require greater effort and delayed rewards.

Keywords: learning strategies; effective learning strategies; instruction; direct experience; durability of intervention

1. Introduction

College students exhibit wide-ranging differences in their ability to effectively learn and retain new material (Cassidy, 2012; Chen et al., 2000). While inherent cognitive abilities account for some of the divergence in learning outcomes (e.g., Lyons & Zelazo, 2011; Ahmed et al., 2019), research has indicated that disparities in learning performance are also related to variations in how students approach learning (Bowen & Wingo, 2012; Bartoszewski & Gurung, 2015; Hassanbeigi et al., 2011; Hartwig & Dunlosky, 2012). Unfortunately, students typically receive little or no formal instruction on the effectiveness of various

learning practices. As such, they are largely left to their own devices with respect to how to approach learning. Research shows that students frequently utilize a host of practices, including highlighting, rereading repeatedly, and cramming (Kornell & Bjork, 2007; Hartwig & Dunlosky, 2012; Karpicke et al., 2009).

Given the diversity of students' approaches, numerous studies have examined the effectiveness of different learning strategies (e.g., Karpicke et al., 2009; Bartoszewski & Gurung, 2015; Hattie & Donoghue, 2018). Highlighting, for instance, is a strategy most students have reported using (Gurung et al., 2010). Thus, many researchers have examined the effect of highlighting/underlining on retention and comprehension (e.g., Hoon, 1974; Ponce et al., 2022). For example, Fowler and Barker (1974) had participants read articles about boredom and city life. Participants were divided into four groups: those who actively highlighted the text, those who passively read the text highlighted by peers, those who read pre-highlighted text by experimenters, and those who read the text without highlighting. One week later, participants took a 54-item multiple-choice test on the material. Participants across the conditions did not significantly differ in their overall test scores, suggesting that, at a minimum, the act of highlighting brought no extra value for learning above and beyond the act of reading (Dunlosky et al., 2013, but also see Ponce et al., 2022 for an alternative view, particularly when highlighting is combined with other practices).

Researchers have likewise examined the impact of numerous learning strategies that students rarely come to utilize on their own. For example, Pan et al. (2025) examined the benefits of interleaved practice in second language (L2) grammar learning by having participants practice verb conjugations and tense identification in Spanish and French using either an interleaved or blocked schedule. Participants in the interleaved condition alternated between tenses during their practice sessions, while participants in the blocked condition practiced one tense at a time. Following the practice sessions, participants' retention and application of the grammar were tested using a grammaticality judgment test where participants evaluated sentences for correct grammar use. The results showed that participants in the interleaved condition outperformed those in the blocked condition on verb conjugation, tense identification, and language identification tasks. Similarly, numerous studies have now demonstrated that active retrieval-based learning strategies (e.g., testing oneself) produce better learning than more passive approaches (e.g., re-reading) (Roediger & Butler, 2011; Roediger & Karpicke, 2006). For instance, Butler and Roediger (2007) demonstrated that active retrieval practice after lectures significantly improved material retention compared to passively reviewing summaries. Likewise, Bartoszewski and Gurung (2015) found that students who employed practice testing and self-explanation achieved significantly better outcomes on exams compared to those who relied on summarizing or highlighting.

Given that there is now a sizable body of research focused on the effectiveness of many different learning strategies, this has enabled the use of systematic review methods to assess the impact of these learning strategies across various contexts. In their systematic review, Dunlosky et al. (2013) revealed that strategies that promote active engagement with the material, such as practice testing, interleaved practice, and distributed practice, consistently enhance comprehension and retention. In contrast, strategies that encourage less active engagement, like summarizing, highlighting, and rereading, often produce limited benefits.

Researchers have also examined whether students are aware of which strategies are the most effective. Studies surveying college students about their learning strategies revealed that students commonly rely on ineffective approaches such as rereading, highlighting and massing practice (Kornell & Bjork, 2007; Karpicke et al., 2009; McCabe, 2011; Witherby et al., 2025). For example, Kornell and Bjork (2007) found that students largely prioritized urgent deadlines, with 59% studying whatever was due soonest (cramming) and only 11%

studying in advance (distributing practice). This result is mirrored by work by Hartwig and Dunlosky (2012), who found that 83.6% of students reported massing their study sessions, and 66% reported rereading (both of which are ineffective strategies). Even when students use strategies like self-testing, many employ self-testing primarily to monitor their understanding rather than enhance their learning.

These findings highlight the need for approaches that could reduce the extent to which students utilize ineffective learning strategies and increase the extent to which they adopt more effective learning strategies. Research has shown that many students engage in less effective practices due to misconceptions about their effectiveness or a lack of awareness of more efficient alternatives (Hartwig & Dunlosky, 2012; Karpicke et al., 2009; McCabe, 2011). Consequently, most studies have focused on addressing these misconceptions through interventions designed to improve students' understanding of evidence-based learning strategies via one of two main approaches.

The first general approach is instruction, where students are explicitly taught about the effectiveness of learning strategies through videos, lectures, or course modules (Ariel & Karpicke, 2018; Cathey et al., 2016). This method aims to enhance students' awareness and understanding of evidence-based practices. Studies of this type have produced somewhat mixed results with respect to their impact on changing students' knowledge and subsequent behaviors (Dembo & Seli, 2004; Foerst et al., 2017; McCabe, 2011). In a more positive direction, Brown-Kramer (2022) had students read one of four empirical articles on learning strategies—distributed practice, rereading, practice testing, or forming mental images—and write a paper to summarize, analyze, and apply the findings. Over the semester, students reported increased use of effective strategies such as practice testing and interleaved practice, while reducing their reliance on less effective strategies like highlighting and summarization. In the less positive direction, Cathey et al. (2016) embedded a midterm study skills training session into an Introductory Psychology course. They presented participants with videos focusing on effective study strategies and common misconceptions about learning. Although students who attended the intervention gained greater insight into their study strategy weaknesses and rated the session as helpful and effective, the training did not significantly change their study time or approach to learning. In sum, it appears that pure instruction-based interventions may produce slight shifts in beliefs, but it remains unclear how much they change learning behaviors, as instruction alone in particular often fails to produce long-term behavioral changes (Balch, 2001; Dembo & Seli, 2004; Foerst et al., 2017).

The second general approach is to provide direct experience by having students actively engage with effective learning strategies. This method aims to have students practice strategies like practice testing and interleaving to help them understand the benefits of these strategies (Einstein et al., 2012). Similarly to the instruction approach, this approach also produced mixed results (DeWinstanley & Bjork, 2004; Gurung & Burns, 2019). For instance, in the positive direction, Balch (2006) provided students with the opportunity to experience distributed versus massed practice of word pairs in a classroom experiment. Students performed better with the distributed pairs, and a follow-up evaluation revealed that they reported a slight shift in their perception and study habits toward using distributed practice. While research indicates that engaging with more effective strategies may encourage positive behavioral changes, individuals often fail to recognize what is beneficial because they misinterpret their experiences. For instance, Pan and Rivers (2023) had participants experience pretesting across multiple cycles and demonstrated that participants had better learning outcomes. However, their perspective of the strategy only changed when they were provided with external guidance that highlighted their improved performance with pretesting. Similarly, Yan et al. (2016) showed that, even after participants benefited

from interleaving during category learning, many still believed blocking to be more effective. Only when participants' attention was explicitly directed to the connection between their performance and the study schedule did their beliefs begin to shift—highlighting the challenge of correcting metacognitive illusions through experience alone.

Both instruction and experience-based approaches have clear benefits but often fail to produce significant changes in understanding and behavior when used alone. Instruction increases awareness but rarely leads to lasting behavioral change, while experience provides opportunities for implementation, but rarely produces the understanding without explicit guidance. Yüksel et al. (2024) combined these two approaches by integrating direct instruction with experience. They took advantage of a natural experimental intervention in a college classroom where one instructor of Introduction to Psychology both taught students about effective strategies in a full dedicated lecture and also embedded these practices into the course structure. For instance, students learned about the benefits of spacing learning sessions over massed practice (i.e., cramming) as part of the lecture on effective learning practices in academia. They also took cumulative quizzes throughout the semester designed to make them space their study sessions (with the instructor clearly tying the practice to the explicit instruction). Yüksel et al. (2024) demonstrated that this combination of direct experience and direct practice produced notable gains in understanding and use of learning strategies in accordance with the empirical evidence around effectiveness.

An open question is whether these improvements persist beyond the intervention. Few studies have examined the long-term durability of these changes over extended periods, such as a semester or longer, to determine whether these strategies continue to influence behavior and performance (e.g., Usdansky et al., 2024; Brown-Kramer, 2022). However, the small number of studies that have done so suggest that producing long-term changes in behavior is likely possible. For instance, Brown-Kramer (2022) examined the long-term effects of an intervention in which students completed term papers on learning strategies. The study compared a control semester to one where students wrote about varied strategies. Results showed that students who engaged with effective strategies, such as practice testing, demonstrated improved strategy use and better course performance in the following semester.

To further assess the extent to which long-term changes in beliefs and use of effective and ineffective learning strategies are feasible, the current research examines the long-term impact of the intervention described above in Yüksel et al. (2024), which utilized both explicit instruction about both effective and ineffective learning strategies as well as direct experience with effective strategies. In Study 1, we assessed whether participants maintained changes in their beliefs about learning strategies and their effectiveness six months after the intervention. In Study 2, we expanded the scope by including students from various academic years by analyzing data from those who enrolled in the intervention section of the course over a five-year period. To foreshadow the results, we found that reported behaviors (i.e., the learning strategies that students actually used) often belied their knowledge. Indeed, students strongly retained their understanding of which approaches were likely to be more or less effective, but nonetheless frequently reported utilizing strategies they knew were less effective. As such, we also gathered information via open-ended questions with the specific goal of understanding the obstacles that the students perceived were preventing them from using more effective learning strategies.

2. Study 1

The current work builds upon a previously published intervention study (Yüksel et al., 2024) in which students in one section of an Introduction to Psychology course received instruction on and experience with effective practices in learning (intervention group),

while students in other sections did not (business-as-usual group). While Yüksel et al. (2024) only examined the impact of the intervention shortly after the conclusion of the intervention, here we collected long-term follow-up data to determine whether the changes in beliefs about, and use of, effective learning strategies that were observed directly post-intervention persisted long-term. Specifically, for Study 1, the new data corresponds to the long-term follow-up assessments collected approximately six months after the intervention. The corresponding subset of pre-intervention and post-intervention data (the full set of which was reported in Yüksel et al., 2024) are included here to provide a clear point of comparison. In the intervention condition, surveys were administered at three time points labeled relative to the intervention period: "pre-intervention" (before the intervention), "post-intervention" (at the end of the course), and "long-term" (approximately six months later). We administered surveys in the business-as-usual condition at times comparable to the administration of surveys in the intervention condition; that is, the pre-intervention survey occurred before the intervention period, the post-intervention survey at the end of the course, and the long-term follow-up six months later.

2.1. Method

2.1.1. Participants

We reached out to the 316 (intervention: 118, business-as-usual: 198) participants who had taken part in the intervention portion of Yüksel et al. (2024). In total, 62 participants agreed to participate, of which 54 (38 women, *Mage* = 19.11) had pre-intervention data available. All participants provided consent in accordance with an approved IRB protocol, and those who participated in the study received a USD 5 Amazon gift card.

2.1.2. Procedure

While the focus of the current manuscript is on the long-term persistence of changes in learning strategy use as a function of the interventions utilized in Yüksel et al. (2024), here we briefly describe the full set of methods (see Yüksel et al., 2024, for a complete description of all procedures).

Intervention: The intervention group (N = 15) was one section of Introduction to Psychology that included a dedicated module aimed at teaching effective learning strategies. This module highlighted evidence-based effective learning practices, such as spacing study sessions over time rather than cramming, interleaving topics instead of focusing on one at a time (blocking) and employing active learning techniques like retrieving information rather than passively reading it. These strategies were taught and framed in terms of their benefits for long-term learning. The instruction was designed to promote active engagement with the material, fostering deep processing over passive learning. In addition to covering effective learning strategies, the module also mentioned commonly used learning strategies with limited effectiveness, such as highlighting text passively, rereading, copying notes, summarizing, cramming, and using keyword mnemonics.

To support students' understanding of these strategies, the course also integrated the general principles into the semester's structure. This included cumulative weekly quizzes to encourage interleaving and spacing, varied seating arrangements to support contextual variability, and activities like generating quiz questions to facilitate deeper processing. Students not only practiced these strategies but also learned the rationale behind them. For instance, they were taught that cumulative exams are designed to promote interleaving and spacing in their study routines.

Business-as-Usual: The business-as-usual sections of Introduction to Psychology also covered some of the materials covered in the intervention section, but in a far less extensive fashion. Similarly, while a few strategies were implemented to some extent

in the business-as-usual sections (e.g., weekly non-cumulative quizzes), these practices were less comprehensive than the intervention section. Because there were no significant differences in either perceived effectiveness or reported frequency of use among any of the business-as-usual courses (N = 39), these were combined into a single group.

2.1.3. Assessments

For each of the 16 different learning strategies—rereading, group studying, cramming, highlighting, using flashcards, thinking of real-life examples, note-taking, practice tests, looking over notes, copying notes, summarizing, creating an outline, studying in different places, spacing, self-explanation, and interleaving—participants were asked to report the extent to which they used the given strategies and whether they viewed the strategies as effective/ineffective. Specifically, participants rated the frequency of use and perceived effectiveness of each of the learning strategies on a 1–10 scale, with 1 indicating "not at all" and 10 indicating "very frequently" or "very effective". For frequency of use, participants responded to the prompt: "How frequently have you been using each of the following studying methods this semester?" For perceived effectiveness, they answered: "How would you rate the effectiveness of each studying method?" (see Appendix C for an example, and the questionnaires folder on OSF for the full survey).

The participants' responses were then matched with their responses taken prior to the intervention (pre-test) and/or directly after the intervention (post-test). Four of the sixteen learning strategies were added for this study and thus have no corresponding pre-test/post-test data (studying in different places, spacing, self-explanation, and interleaving). These four strategies were assessed via scenario-based questions in Yüksel et al. (2024); however, to reduce participant burden in the current work, these were switched to simple frequency/effectiveness measures here.

We also included several open-ended questions to gain deeper insights into participants' study behaviors and potential barriers to adopting effective strategies. These questions asked participants to list the classes they were taking that semester (e.g., "Please list all of the classes that you're taking this semester"), describe their study habits, indicate any changes in study habits compared to the previous semester, and identify strategies they had not used but would like to. Additionally, we asked them to reflect on the barriers preventing them from using the strategies they perceive as effective ones (e.g., "What are the barriers to using study habits you have not used this semester but would like to use?").

All assessments were conducted online via Qualtrics. During the pre-test and post-test phases, participants completed the measures in the following order: perceived effectiveness, reported frequency of use, and scenario-based questions. The items and scenarios were presented in a randomized order. However, the scenarios are not relevant to the current study, as they were not included in the long-term follow-up. In the long-term follow-up, participants completed open-ended questions first, followed by the perceived effectiveness and reported frequency of use measures. The order of items within each measure was randomized.

2.2. Results

All data analyses were performed through R (R Core Team, 2017). The data files and script folders, where all data analyses are performed, are available on the Open Science Framework (OSF, https://osf.io/kcduz/?view_only=8ba58424b68743c9a39a6e7b66c34516 accessed on 24 April 2025).

The dependent variables were the perceived effectiveness and reported frequency of use of effective and ineffective strategies. For the main analyses of perceived effectiveness and reported frequency of use, we calculated two separate aggregate scores by grouping

learning strategies into effective and ineffective learning strategies. We followed the classification from Dunlosky et al.'s (2013) systematic review and the related literature to define these categories. Effective strategies included taking tests, using flashcards, applying real-life examples, creating outlines, group study, studying in different places, spacing the study schedules, self-explanation, and interleaving the study material. While such a binary (effective/ineffective) classification scheme is not capable of capturing various nuances (e.g., the extent to which certain strategies are effective can depend on the manner in which they are employed), it was utilized in the current work for several reasons. First, we used the classification from Dunlosky et al. (2013) as a framework for presenting learning strategies to students. Indeed, this categorization allowed for a method that was independent of the research team. That is, what was described as effective was based not on instructor's beliefs, but rather the science of learning. Finally, this categorization also allows for a direct match with previously published work using pre-registered analyses (Yüksel et al., 2024; note the current work was not pre-registered, but did directly follow the previous work that was). For readers interested in possible alternative schemes the data fully broken out by individual strategies are available in the Appendix A.1.

Because we included four new effective learning strategies in this follow-up work, we calculated two scores for effective learning strategies: one for the combination of the 12 learning strategies utilized in Yüksel et al. (2024) and one for the full set of 16 strategies included in the current study. Then, we confirmed that these two measures were not significantly different from each other, either for perceived effectiveness ($t(53) = 1.55$, $p = 0.13$, 95% CI [−0.05, 0.36]) or for reported frequency of use ratings ($t(53) = -0.33$, $p = 0.74$, 95% CI [−0.23, 0.16]). Ineffective strategies consisted of highlighting, copying notes, copying the textbook, cramming, rereading, and summarizing. Higher scores reflected greater understanding, perceived effectiveness, and frequency of use.

2.2.1. Pre-Intervention Comparison of Groups

We first assessed whether there were any significant differences between the intervention group and the business-as-usual group prior to our analyses to ensure that any observed differences emerged after the intervention. At pre-intervention, participants did not differ significantly on any outcome: perceived effectiveness of effective, $F(1,43) = 1.19$, $p = 0.28$, or ineffective strategies, $F(1,43) = 0.92$, $p = 0.34$; reported frequency of use of effective, $F(1,43) = 0.73$, $p = 0.78$, and ineffective strategies, $F(1,43) = 0.36$, $p = 0.54$. Thus, there is no evidence to suggest that the groups differed before beginning the study.

2.2.2. Use and Perceived Effectiveness of Effective Learning Strategies

To determine whether enrollment in the intervention group was associated with perceived effectiveness and use of learning strategies compared to the business-as-usual group, we fitted a series of linear mixed-effects models on four dependent variables: perceived effectiveness of effective strategies, perceived effectiveness of ineffective strategies, reported frequency of use of effective strategies, and reported frequency of use of ineffective strategies. We regressed the dependent variables of interests on group (contrast coded: "−0.5" for business-as-usual and "0.5" for intervention), time (treated as a categorical factor: pre-intervention vs. post-intervention, vs. long-term follow-up) and their interaction, controlling for the random effects of time.

Perceived Effectiveness of Learning Strategies. Two linear mixed effect models were conducted to examine the effects of time, group, and their interaction on the perceived effectiveness of effective and ineffective learning strategies, controlling for random effects of time.

For effective learning strategies, the results revealed no significant main effect of time. Perceived effectiveness did not differ significantly between the post-intervention ($b = 0.29$, $SE = 0.41$, $t = 0.70$, $p = 0.48$) or the long-term follow-up ($b = -0.03$, $SE = 0.41$, $t = -0.07$, $p = 0.95$) compared to the pre-intervention phase. Similarly, the main effect of group, intervention or business-as-usual, was not significant ($b = -0.59$, $SE = 0.50$, $t = -1.18$, $p = 0.24$). An interaction between time and group was significant: students in the intervention group ($M = 7.86$) reported higher perceived effectiveness for effective learning strategies than those in the business-as-usual group ($M = 7.14$) during the long-term follow-up ($b = 1.31$, $SE = 0.67$, $t = 1.98$, $p = 0.04$), whereas the groups did not significantly differ during post-intervention (*Mintervention* = 8.05, *Mbusiness-as-usual* = 7.59) ($b = 1.05$, $SE = 0.67$, $t = 1.58$, $p = 0.117$) compared to the pre-intervention (*Mintervention* = 7.24, *Mbusiness-as-usual* = 7.83) (Table 1, Figure 1a).

Table 1. Mean (standard deviation) of perceived effectiveness ratings for effective and ineffective learning strategies across groups and time points in Study 1.

Perceived Effectiveness						
	Effective Strategies			Ineffective Strategies		
Groups	Pre-Intervention	Post-Intervention	Long-Term Follow-Up	Pre-Intervention	Post-Intervention	Long-Term Follow-Up
Business-as-usual	7.83 (1.44)	7.59 (1.45)	7.14 (1.39)	6.80 (1.42)	5.97 (1.41)	5.96 (1.32)
Intervention	7.24 (1.92)	8.05 (1.36)	7.87 (1.25)	6.31 (1.63)	4.28 (1.44)	4.49 (1.16)

For ineffective learning strategies, the main effect of time was not significant; however, we observed a trend. Students reported lower perceived effectiveness of ineffective strategies during the post-intervention ($M = 5.50$) ($b = -1.43$, $SE = 0.76$, $t = -1.87$, $p = 0.06$) and long-term follow-up ($M = 5.55$) ($b = -1.33$, $SE = 0.76$, $t = -1.74$, $p = 0.08$) compared to the pre-intervention ($M = 6.68$). The main effect of the group was not significant ($b = -0.49$, $SE = 0.48$, $t = -1.01$, $p = 0.312$). The interaction between time and group was also not significant. Overall, students in the intervention group ($M = 4.28$) reported lower perceived effectiveness than those in the business-as-usual group ($M = 5.97$) during post-intervention ($b = -1.20$, $SE = 0.64$, $t = -1.88$, $p = 0.062$), whereas the groups made similar ratings in the pre-intervention phase (*Mintervention* = 6.31, *Mbusiness-as-usual* = 6.79) and the long-term follow-up (*Mintervention* = 4.48, *Mbusiness-as-usual* = 5.95) (Table 2, Figure 1b).

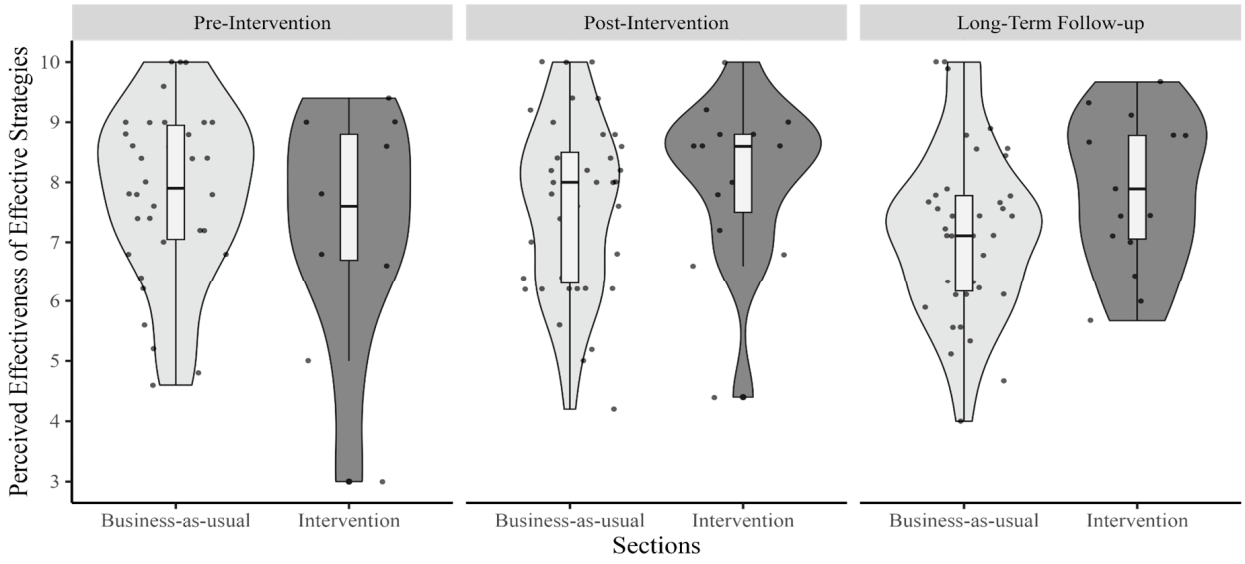
Table 2. Mean (standard deviation) of reported frequency of use ratings for effective and ineffective learning strategies across groups and time points in Study 1.

Reported Frequency of Use						
	Effective Strategies			Ineffective Strategies		
Groups	Pre-Intervention	Post-Intervention	Long-Term Follow-Up	Pre-Intervention	Post-Intervention	Long-Term Follow-Up
Business-as-usual	6.68 (1.27)	5.33 (1.76)	5.39 (1.48)	6.40 (1.37)	5.74 (1.55)	5.78 (1.21)
Intervention	6.55 (1.97)	5.73 (1.36)	6.34 (1.13)	6.70 (1.73)	4.74 (1.29)	5.39 (1.35)

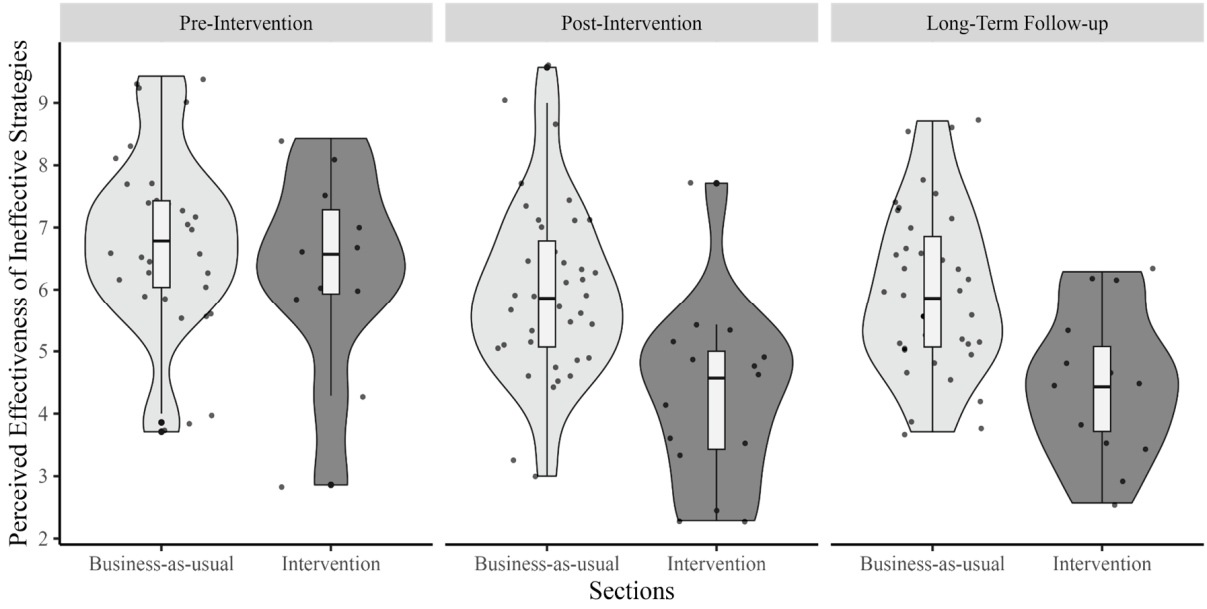
Reported Frequency of Use of Learning Strategies. As above, two linear mixed effect models were constructed to examine the effects of time, group, and their interaction on the reported frequency of use of effective and ineffective learning strategies.

For effective learning strategies, the results revealed a significant main effect of time. Participants reported significantly lower frequency of effective strategy use in the post-

intervention (M = 5.44) (b = −1.08, SE = 0.35, t = −3.11, p = 0.004) and in the long-term follow-up phase (M = 5.65) (b = −0.75, SE = 0.35, t = −2.15, p = 0.03) compared to the pre-intervention (M = 6.65). We did not find a significant main effect of group (b = −0.14, SE = 0.52, t = −0.26, p = 0.80). Moreover, the interaction between time and group was not significant such that students' ratings did not differ across groups during the post-intervention (b = 0.54, SE = 0.70, t = 0.78, p = 0.44) or long-term follow-up (b = 1.09, SE = 0.70, t = 1.56, p = 0.12) compared to the pre-intervention (Figure 2a).



(**a**)

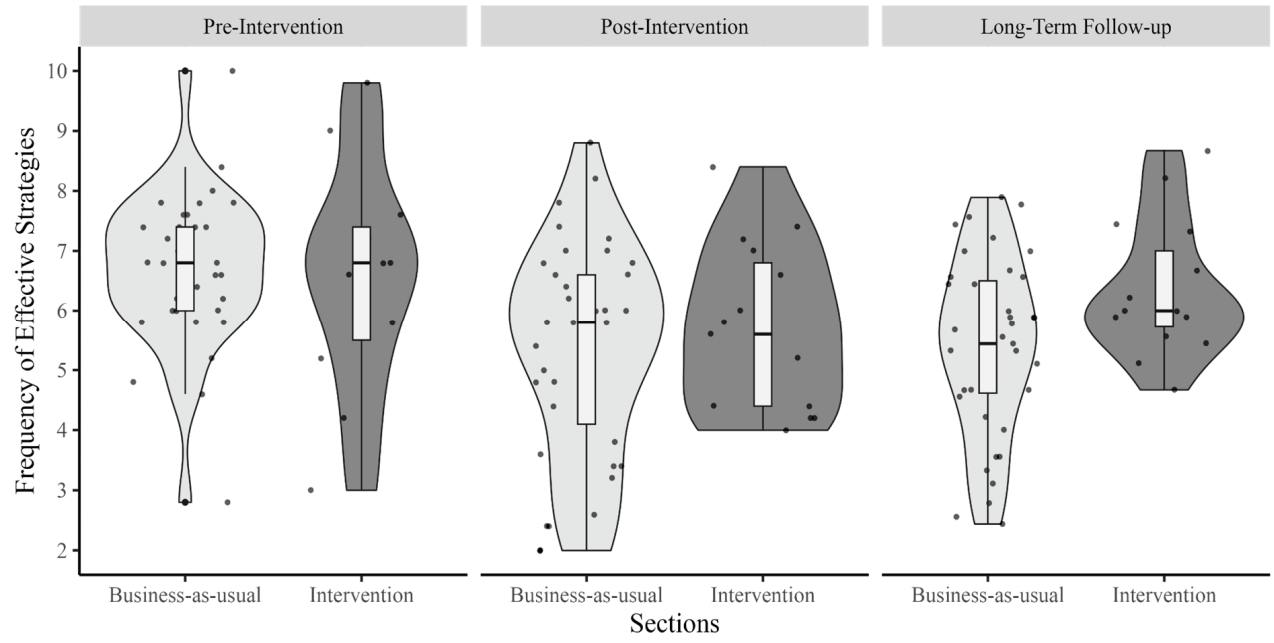


(**b**)

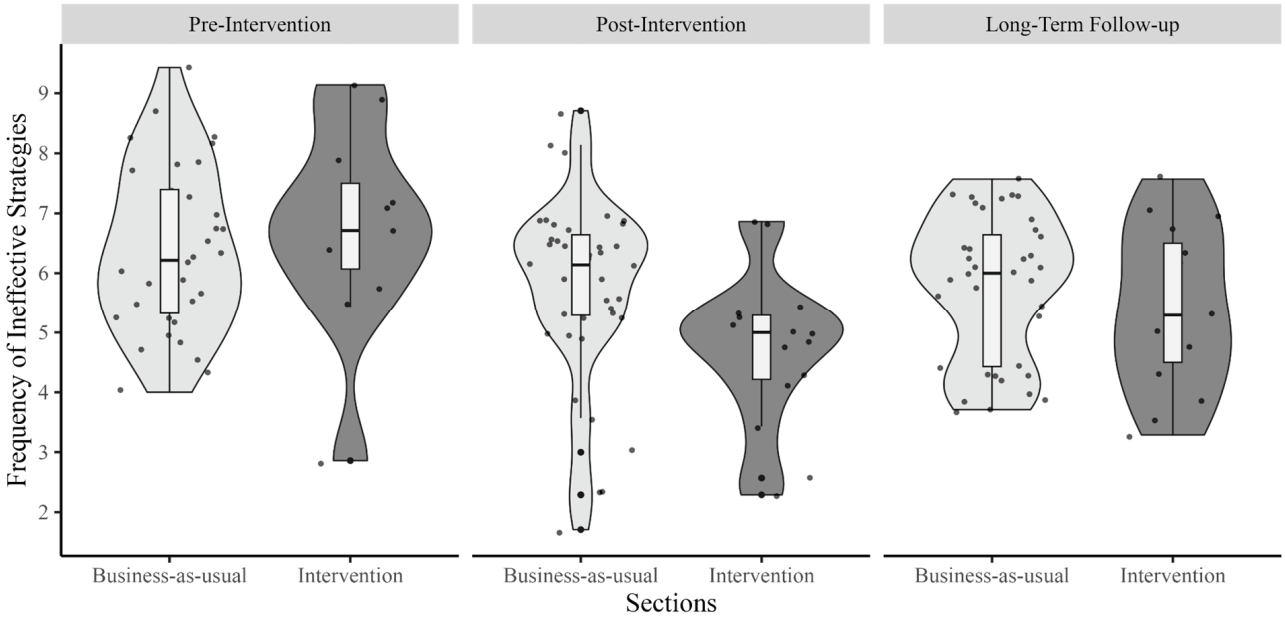
Figure 1. (**a**) Pre-intervention, post-intervention, and follow-up ratings of perceived effectiveness of effective strategies across groups. (**b**) Pre-intervention, post-intervention, and follow-up ratings of perceived effectiveness of ineffective strategies across groups.

Similarly, for ineffective learning strategies, there was no significant difference in students' ratings across time points. However, a trend was observed, such that students' ratings for frequency of ineffective strategy use were lower in the post-intervention (M = 5.46) (b = −1.31, SE = 0.67, t = −1.93, p = 0.055) and in the long-term follow-up (M = 5.67)

($b = -0.96$, $SE = 0.67$, $t = -1.43$, $p = 0.15$) compared to their ratings in the pre-intervention ($M = 6.47$). No significant main effect of group was found ($b = 0.31$, $SE = 0.49$, $t = 0.63$, $p = 0.53$). There was an interaction effect between time and group: students in the intervention group reported less frequency of use of ineffective learning strategies during the post-intervention compared to the pre-intervention ($b = -1.30$, $SE = 0.64$, $t = -2.02$, $p = 0.04$); however, their ratings were not significantly different across the long-term follow-up ($b = -0.69$, $SE = 0.64$, $t = -1.07$, $p = 0.28$) and the pre-intervention (Figure 2b).



(**a**)



(**b**)

Figure 2. (**a**) Pre-intervention, post-intervention, and follow-up ratings of frequency of use of effective learning strategies across groups. (**b**) Pre-intervention, post-intervention, and follow-up ratings of frequency of use of ineffective learning strategies across groups.

While we calculated aggregate scores for effective and ineffective strategies to facilitate interpretation in our main analyses, we also included detailed analyses of each individual learning strategy in Appendix A.1, Tables A1 and A2. Additionally, due to the low number of responses to open-ended questions, we report them descriptively in Appendix A.2, Tables A3 and A4.

3. Study 2

In Study 1, we observed that some of the changes in understanding about effective/ineffective learning strategies observed immediately after the intervention persisted six months after the intervention. However, there was no sustained change in the reported frequency of use of learning strategies. Study 2 was conducted to further investigate whether the effects on understanding remained consistent over an extended period and whether they eventually contributed to any meaningful changes in the use of effective or ineffective strategies. Moreover, we aimed to expand our participant pool in Study 2; we reached out to all students who had enrolled in the Introduction to Psychology course at University of Wisconsin-Madison during the past five years. By gathering responses from students across different sections and academic years, we could examine if there were any differences in understanding or use of learning strategies across sections that might emerge over time.

3.1. Method

3.1.1. Participants

We contacted a total of 12,611 (intervention: 1454, business-as-usual: 11,156) students who enrolled in various sections of Introduction to Psychology at University of Wisconsin-Madison over the past five years. In total, 117 (90 women, *Mage* = 20.87) participants opted to take part in the study and were entered into a USD 100 lottery for their participation. Among these participants, 102 (77 women, *Mage* = 20.84) completed the full survey and passed the attention check question that was asked to see if participants were properly reading all the questions in the survey. Data from these 102 participants were used in the quantitative analyses. Of those, 68 were previously enrolled in one of the business-as-usual sections and 34 were enrolled in the intervention section. However, for the open-ended questions, we included responses from all 117 (80 business-as-usual, 37 intervention) participants who answered, regardless of whether they completed the full survey or passed the attention check, in order to retain all available qualitative data. Notably, nine of these participants were enrolled in the same semester as the participants in Study 1 (Spring 2021, see Appendix B, Table A5 for full list), and only one participant took part in both studies (the inclusion/exclusion of this participant made no difference in the reported outcome and thus they were included to maximize the size of the dataset).

3.1.2. Assessments

Participants were asked to complete the same set of questionnaires used in Study 1, which included ratings of the perceived effectiveness and reported frequency of use of both ineffective and effective learning strategies. All assessments were administered post-intervention. Differently from Study 1, Study 2 also included scenario-based questions from Yüksel et al. (2024) that were designed to assess participants' ability to apply learning strategies in context (see Appendix C, Figures A1 and A2 for examples and OSF for the full set). Additionally, participants answered open-ended questions about their general study behaviors and described any obstacles they encountered when attempting to use effective strategies.

Participants responded to a series of open-ended questions designed to assess their study habits, learning strategies, understanding of learning strategies, and perceived barriers. These questions were adapted from Study 1, but this study aimed to examine these areas in greater detail. Participants were asked to describe their current study routines and habits for their classes and exams, assess whether their study habits and learning strategies were effective for their learning, and provide explanations for their evaluations (e.g., Do you think that the study habits/learning strategies you are using are the effective ones for your learning needs? Why/not?). They also reflected on how their study routines and habits had changed over time in college and listed any new study habits or learning strategies they were using (e.g., How have your study routines/habits changed over the course of your time in college?). Additionally, participants identified study habits or learning strategies they had not yet implemented but would like to, along with the barriers preventing them from doing so (e.g., Are there any study habits/learning strategies that you have not used, but would like to implement? Please list all of the habits/strategies that you can think of; What are the barriers to using study habits/learning strategies you have not used this semester, but you would like to use?). Finally, they elaborated on the obstacles preventing them from adopting strategies they had heard or learned were effective but had not yet utilized (e.g., What are the barriers preventing you from using study habits/learning strategies that you have heard/learned are effective but have not yet used?).

We also included the list of scenarios utilized in Yüksel et al. (2024) which served as a measure of participants' understanding of learning strategies. These scenarios were included in the current study to evaluate whether, after a significant amount of time, participants could identify effective and ineffective strategies in hypothetical contexts. Participants were presented with various scenarios and asked to choose which strategy they would most likely use if they were the student in the given situation. For instance, participants were presented with a scenario in which a professor offered two study options to help students learn painting styles by three different artists: one slideshow grouped all paintings by artist, while the other alternated paintings between artists. Participants were asked to choose which slideshow they would prefer to use if they were a student in this class (see more examples from https://osf.io/kcduz/?view_only=8ba58424b68743c9a39a6e7b66c34516 accessed on 24 April 2025).

3.2. Results

Consistent with Study 1, we calculated aggregate scores for the perceived effectiveness and reported frequency of use of both effective and ineffective learning strategies. Additionally, we calculated aggregate scores for the scenario questions. For effective learning strategies and scenario outcomes, higher scores reflected a better understanding of effective learning strategies. Conversely, for ineffective learning strategies, lower scores indicated a better understanding, as they demonstrated the ability to recognize and avoid less effective strategies. Following this, we conducted several linear regression analyses where the dependent variables were regressed on the groups (intervention vs. business-as-usual), controlling for the semester (from 2020 Fall to 2024 Spring). We included semester as a control variable rather than conducting semester-level analyses because the number of participants per semester was limited (e.g., some semesters had no participants; see Appendix B, Table A5 for full breakdown of participants by group and semester).

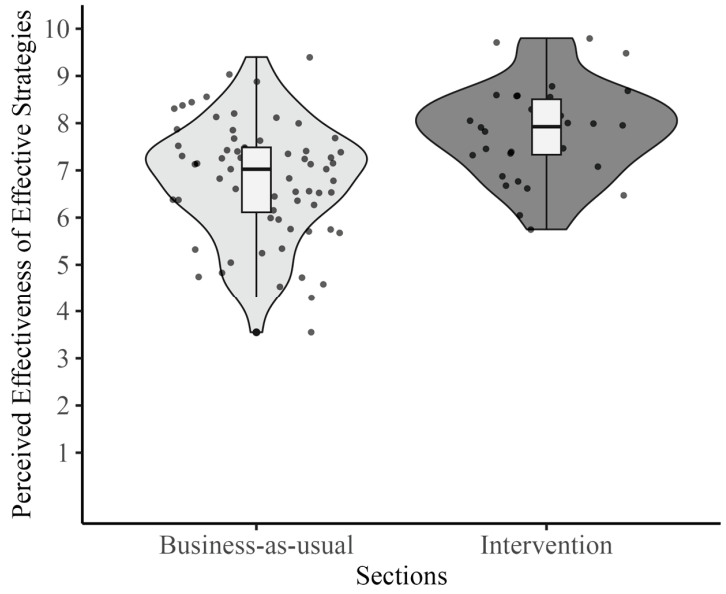
3.2.1. Quantitative Assessments

Perceived Effectiveness of Learning Strategies. The analysis revealed that the perceived effectiveness of effective strategies was significantly higher for the intervention

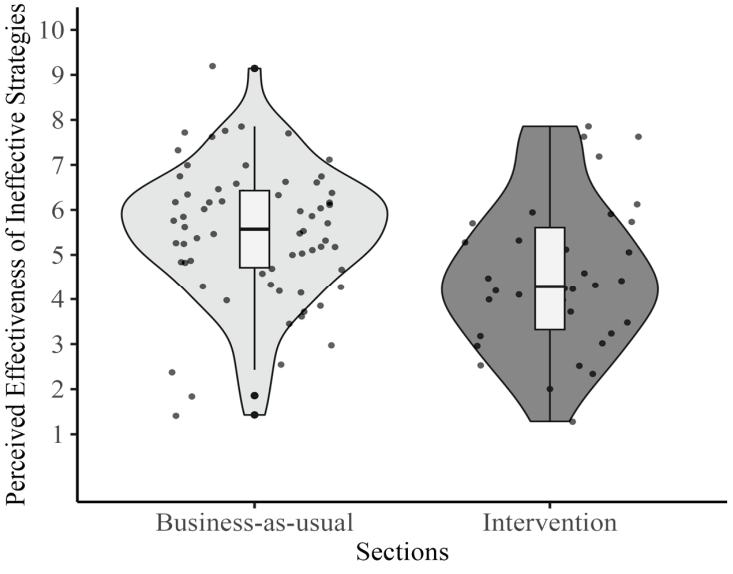
group, $b = 1.36$, $t(91) = 4.38$, $p < 0.001$, with means of 7.83 and 6.79, respectively (see Table 3, Figure 3a). In contrast, while the effect on ineffective strategies was numerically the same magnitude as for effective strategies (intervention: $M = 4.50$; business-as-usual: $M = 5.49$), this did not reach significance, $b = -0.46$, $t(91) = -1.15$, $p = 0.25$ (see Table 3, Figure 3b).

Table 3. Mean (standard deviation) of perceived effectiveness ratings for effective and ineffective learning strategies across groups and time points in Study 2.

	Perceived Effectiveness		Reported Frequency of Use		Scenarios
Groups	**Effective Strategies**	**Ineffective Strategies**	**Effective Strategies**	**Ineffective Strategies**	
Business-as-usual	6.79 (1.22)	5.48 (1.47)	5.62 (1.38)	5.60 (1.34)	6.80 (1.02)
Intervention	7.83 (0.96)	4.50 (1.63)	6.51 (1.29)	5.04 (1.52)	8.28 (1.18)



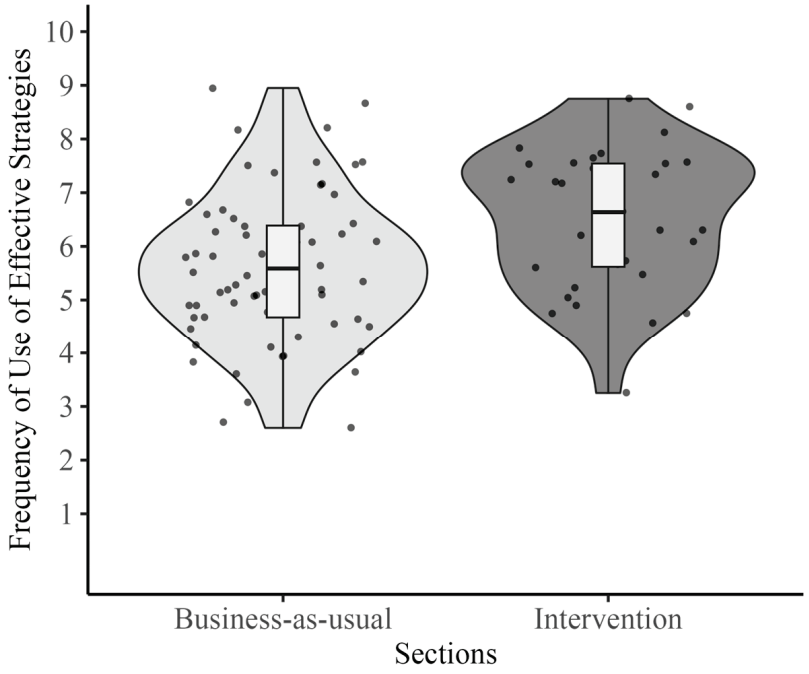
(**a**)



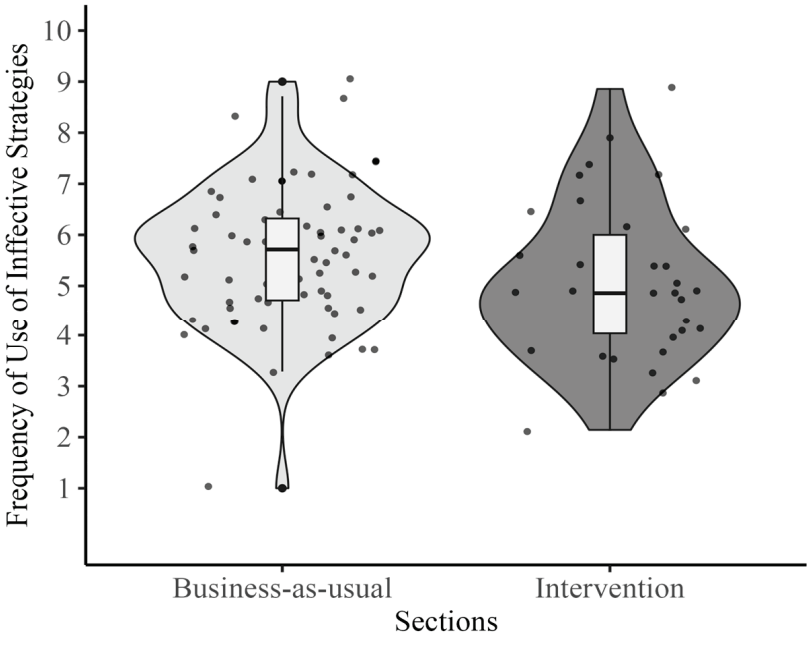
(**b**)

Figure 3. (**a**) Ratings of perceived effectiveness of effective strategies. (**b**) Ratings of perceived effectiveness of ineffective strategies.

Reported Frequency of Use of Learning Strategies. The results demonstrated that there was a significant difference between groups for the reported frequency of effective strategy use, $b = 0.95$, $t(91) = 2.57$, $p = 0.01$, such that students in the intervention group ($M = 6.51$) reported higher ratings compared to those in the business-as-usual group ($M = 5.62$) (see Figure 4a). However, the frequency of using ineffective strategies showed no significant difference between the groups (*Mintervention* = 5.04, *Mbusiness-as-usual* = 5.61), $b = -0.13$, $t(91) = -0.34$, $p = 0.73$ (see Figure 4b).



(**a**)



(**b**)

Figure 4. (**a**) Ratings of reported frequency of use of effective strategies. (**b**) Ratings of reported frequency of use of ineffective strategies.

Scenarios. A linear regression analysis revealed a significant difference between groups in scenario scores, $b = 1.56$, $t(91) = 5.52$, $p < 0.001$. Students in the intervention group had higher scores ($M = 8.28$) than those in the business-as-usual group ($M = 6.81$) (see Figure 5).

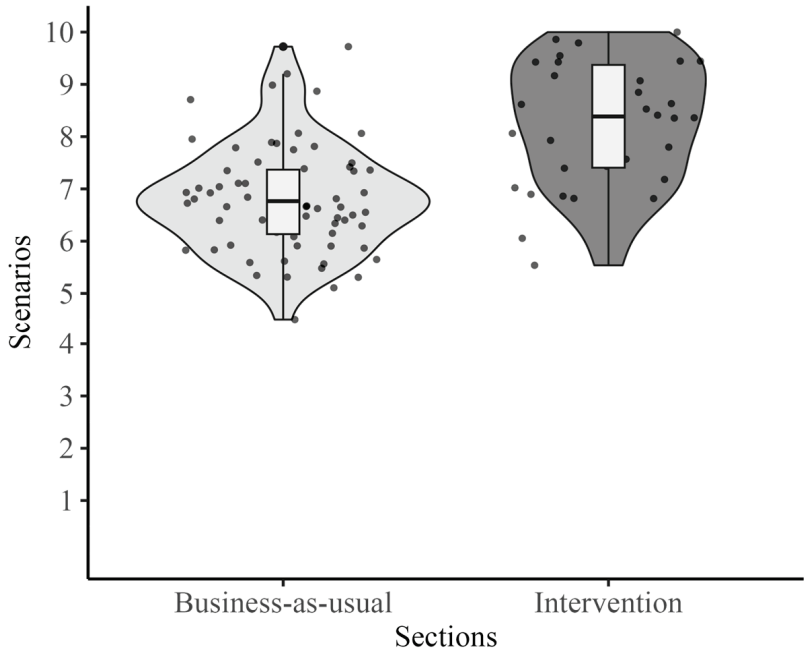


Figure 5. Scenarios ratings.

Although our main analyses relied on aggregate scores for effective and ineffective strategies, we also provided item-level analyses for each individual learning strategy in Appendix B.2, Tables A6 and A7.

3.2.2. Open-Ended Questions

Students' responses to open-ended questions were coded by two independent coders who were unaware of the study's predictions or participants' groups (intervention vs. business-as-usual). We created three main coding schemes including learning strategies mentioned, reasons strategies were perceived as effective or ineffective, and obstacles (see OSF link above for the coding schemes). The coding schemes were developed collaboratively by the first independent coder (outside of the research team, but aware of the general research questions) and the experimenter (E.Y.) through discussions of approximately 15 responses for each question. Subsequently, both coders applied the finalized coding scheme to the entire dataset. Given the fully open-ended nature of the responses, all relevant matches (e.g., learning strategies, obstacles) were coded. Any discrepancy in the listed responses between coders was counted as a mismatch. Using this method, an inter-rater reliability of 87% was achieved.

The main disagreement was related to the definition of cramming versus spacing as a learning strategy. One coder initially categorized studying 4–7 days before the exam as studying in advance (and thus a "spacing" strategy), whereas the other coded this as cramming. After discussion, it was decided to set the cutoff for defining cramming as less than 7 days before an exam. After finalizing the coding based on the final discussion, the number of participants in each condition for each response was counted. The number of participants (and percentages) reported below reflect only those who responded to each specific question. Percentages are calculated by dividing the number of participants who

mentioned a given strategy, reason, or obstacle by the total number of participants who answered that question. Because some participants skipped certain questions (e.g., did not report obstacles but explained why their strategies were effective), the denominators vary. Participants could be counted in multiple categories if they mentioned more than one strategy, reason, or obstacle. For instance, if a participant stated that they regularly study for their exams daily or weekly and also study for long hours the night before an exam, they would be coded both as using spacing (i.e., distributed practice) and as cramming.

Can You Describe Your Current Study Routine and Habits for Your Current Classes and Exams? Students provided detailed information about their current study routines (see Table 4). They listed many learning strategies within the spectrum of both effective and ineffective approaches. They also reported several learning strategies that have not been extensively studied in the literature.

Table 4. Frequency and percentage of reported learning strategies by group.

Response	Business-as-Usual	Intervention	Total
Retrieval Practice	37/76 (49%)	15/35 (43%)	52/111 (47%)
Rereading	33/76 (43%)	14/35 (40%)	47/111 (42%)
Cramming	21/76 (28%)	8/35 (23%)	29/111 (26%)
Time Management	22/76 (29%)	7/35 (20%)	29/111 (26%)
Spacing	20/76 (26%)	8/35 (23%)	28/111 (25%)
Using Flashcards	22/76 (29%)	4/35 (11%)	26/111 (23%)
Studying in Advance	6/76 (8%)	9/35 (26%)	15/111 (14%)
Copying Notes	8/76 (11%)	4/35 (11%)	12/111 (11%)
Creating Outline	8/76 (11%)	0/35 (0)	8/111 (7%)
Group Study	4/76 (5%)	3/35 (9%)	7/111 (6%)
Active Notetaking	2/76 (3%)	4/35 (11%)	6/111 (5%)
Metacognitive Monitoring	3/76 (4%)	3/35 (9%)	6/111 (5%)
Watching Videos	4/76 (5%)	2/35 (6%)	6/111 (5%)
Self-Explanation	3/76 (4%)	1/35 (3%)	4/111 (4%)
Summarization	3/76 (4%)	1/35 (3%)	4/111 (4%)
Using Planner	2/76 (3%)	1/35 (3%)	3/111 (3%)
Visualization	2/76 (3%)	1/35 (3%)	3/111 (3%)
Environment Optimization	1/76 (1%)	1/35 (3%)	2/111 (2%)
Error Analysis	2/76 (3%)	0/35 (0)	2/111 (2%)
Attending Office Hours	1/76 (1%)	1/35 (3%)	2/111 (2%)
Studying in Different Places	0/76 (0)	2/35 (6%)	2/111 (2%)
Highlighting	1/76 (1%)	0/35 (0)	1/111 (1%)
Learning From Examples	1/76 (1%)	0/35 (0)	1/111 (1%)
Listening Audiobooks	1/76 (1%)	0/35 (0)	1/111 (1%)
Interleaved Practice	0/76 (0)	1/35 (3%)	1/111 (1%)

Among the more commonly endorsed effective strategies were retrieval practice and spacing. However, most of the reported spacing was due to weekly quizzes and assignments rather than intentionally spacing study sessions. Among the less active strategies, rereading learning materials—such as textbooks, notes, and slides—was the most commonly reported strategy among both groups (intervention and business as usual). Cramming, defined as heavy studying immediately (~7 days) before exams, was also frequently mentioned by students in both the business-as-usual and intervention groups. Two other less active learning strategies mentioned by students were rewriting notes and creating outlines of the key points. Additionally, a notable number of students expressed efforts to improve their time management (e.g., use of planning tools like Pomodoro) to organize their study schedules. Some participants described learning strategies that have

not been widely studied. For example, listening to audiobooks or watching recorded class videos.

Do You Think That the Study Habits/Strategies You Are Using Are Effective? When students were asked to judge if their learning strategies were effective, the majority of the students (77%) reported that they believed their learning strategies were effective, whereas only a handful of students were unsure, and only a few believed they used ineffective approaches. These ratings were consistent across the groups (see Table 5).

Table 5. Frequency and percentage of students' perceptions of the effectiveness of their own learning strategies: are they effective? (yes, unsure, no).

Response	Business-as-Usual	Intervention	Total
Yes	58/80 (73%)	28/37 (76%)	86/117 (74%)
Unsure	15/80 (19%)	4/37 (11%)	19/117 (16%)
No	7/80 (9%)	5/37 (14%)	12/117 (10%)

Among those who believed their strategies were effective, the majority reported rereading, cramming, and copying notes—approaches that are generally less aligned with long-term learning goals. In contrast, a number of students mentioned strategies like retrieval practice and spacing, which are supported by studies for promoting long-term retention and understanding.

Why/Why Not (In)Effective? As a follow-up to students' judgments of the effectiveness of their learning strategies, they were asked to elaborate on their reasoning. Those who rated their strategies as effective were asked why they found them effective, while those who rated them as ineffective were asked why they believed the strategies were ineffective. Students who were unsure were asked both questions to investigate which aspects they considered effective and which they perceived as ineffective.

When asked why their strategies were ineffective, some students stated that they were aware of more effective alternatives but did not elaborate on why they believed their own strategies were ineffective. A few students attributed their continued reliance on these strategies to procrastination and a lack of organization, while others cited limited access to information about better approaches (see Table 6).

Table 6. Frequency and percentage of reported reasons for perceived strategy ineffectiveness.

Response	Business-as-Usual	Intervention	Total
There Are More Effective Strategies	11/17 (64.7%)	2/6 (33.3%)	13/23 (56.5%)
Procrastination	3/17 (17.6%)	3/6 (50.0%)	6/23 (26.1%)
Not Organized	4/17 (23.5%)	1/6 (16.7%)	5/23 (21.7%)
Lack of Information	3/17 (17.6%)	0/6 (0.0%)	3/23 (13.0%)

When asked why their strategies were effective, achieving good grades was the most frequent reason among both groups. Additionally, the majority of students stated that they believe their methods helped them learn effectively, while some of these students specifically emphasized that their strategies improved understanding rather than memorization. Also, some mentioned their strategies provide a good representation of exams and help them save time (see Table 7).

Changes in Study Routines Over Time. To capture changes in students' approaches over time, two questions were utilized: "How have your study routines/habits changed over the course of your time in college?" and "What are the new study habits/learning

strategies you are using?" Because students frequently responded similarly to both questions, we combined their responses (see Table 8).

Table 7. Frequency and percentage of reported reasons for perceived strategy effectiveness.

Response	Business-as-Usual	Intervention	Total
Good Grades	22/65 (33.8%)	10/32 (31.3%)	32/97 (33.0%)
Learn Effectively	17/65 (26.2%)	14/32 (43.8%)	31/97 (32.0%)
Understand, Not Memorize	12/65 (18.5%)	4/32 (12.5%)	16/97 (16.5%)
Good Representation of Exams	7/65 (10.8%)	2/32 (6.3%)	9/97 (9.3%)
Helps with Thinking Through Problems	4/65 (6.2%)	4/32 (12.5%)	8/97 (8.2%)
Save Time	5/65 (7.7%)	3/32 (9.4%)	8/97 (8.2%)
Feels Productive	5/65 (7.7%)	2/32 (6.3%)	7/97 (7.2%)
Stay Organized	5/65 (7.7%)	0/32 (0.0%)	5/97 (5.2%)
Help Fix Weaknesses	0/65 (0.0%)	2/32 (6.3%)	2/97 (2.1%)

Table 8. Frequency and percentage of reported intended changes to learning strategies.

Response	Business-as-Usual	Intervention	Total
Retrieval Practice	30/76 (39.5%)	17/36 (47.2%)	47/112 (42.0%)
Time Management	23/76 (30.3%)	11/36 (30.6%)	34/112 (30.4%)
Using Flashcards	25/76 (32.9%)	7/36 (19.4%)	32/112 (28.6%)
Rereading	20/76 (26.3%)	10/36 (27.8%)	30/112 (26.8%)
Spacing	12/76 (15.8%)	8/36 (22.2%)	20/112 (17.9%)
Group Study	13/76 (17.1%)	5/36 (13.9%)	18/112 (16.1%)
Copying Notes	13/76 (17.1%)	4/36 (11.1%)	17/112 (15.2%)
Study More	10/76 (13.2%)	4/36 (11.1%)	14/112 (12.5%)
Studying Different Places	5/76 (6.6%)	6/36 (16.7%)	11/112 (9.8%)
Studying in Advance	8/76 (10.5%)	3/36 (8.3%)	11/112 (9.8%)
Environment Optimization	6/76 (7.9%)	4/36 (11.1%)	10/112 (8.9%)
Creating Outline	7/76 (9.2%)	2/36 (5.6%)	9/112 (8.0%)
Metacognitive Monitoring	7/76 (9.2%)	2/36 (5.6%)	9/112 (8.0%)
Visualization	6/76 (7.9%)	3/36 (8.3%)	9/112 (8.0%)
Self-Explanation	3/76 (3.9%)	5/36 (13.9%)	8/112 (7.1%)
Active Note-Taking	2/76 (2.6%)	5/36 (13.9%)	7/112 (6.3%)
Office Hours	7/76 (9.2%)	0/36 (0.0%)	7/112 (6.3%)
Summarization	6/76 (7.9%)	1/36 (2.8%)	7/112 (6.3%)
Using Planner	6/76 (7.9%)	1/36 (2.8%)	7/112 (6.3%)
Watching Videos	5/76 (6.6%)	1/36 (2.8%)	6/112 (5.4%)
Cramming	4/76 (5.3%)	0/36 (0.0%)	4/112 (3.6%)
Interleaved Practice	1/76 (1.3%)	3/36 (8.3%)	4/112 (3.6%)
Highlighting	3/76 (3.9%)	0/36 (0.0%)	3/112 (2.7%)
Real-Life Examples	3/76 (3.9%)	0/36 (0.0%)	3/112 (2.7%)
Mnemonics	1/76 (1.3%)	1/36 (2.8%)	2/112 (1.8%)
Listening to Audiobooks	1/76 (1.3%)	0/36 (0.0%)	1/112 (0.9%)
Elaborative Interrogation	0/76 (0.0%)	1/36 (2.8%)	1/112 (0.9%)
Error Analysis	0/76 (0.0%)	1/36 (2.8%)	1/112 (0.9%)
Learning from Examples	0/76 (0.0%)	1/36 (2.8%)	1/112 (0.9%)

Many students reported an increased use of retrieval practice as a learning strategy over time, a trend observed consistently across both groups. Additionally, improvements in time management were commonly mentioned, with several students indicating that they had developed better organizational skills and planned to continue working on their time management skills. Rereading learning materials, such as textbooks and notes, was another

strategy that students reported both using and intending to use more frequently. Less commonly mentioned changes included strategies such as spacing and creating outlines of key points in the learning material. Interestingly, some strategies were only mentioned by students in the intervention group, including interleaved practice and self-explanation.

Obstacles to Adopting New Strategies. To understand the obstacles students face in adopting effective study strategies, two questions were asked: "What are the barriers to using study habits/learning strategies you have not used this semester but would like to use?" and "What are the barriers preventing you from using study habits/learning strategies that you have heard/learned are effective but have not yet used?" Because students provided similar responses to both questions, their responses to these two questions were combined to summarize the obstacles they reported (see Table 9).

Table 9. Frequency and percentage of reported obstacles to implementing learning strategies.

Response	Business-as-Usual	Intervention	Total
Not Enough Time	31/67 (46.3%)	17/30 (56.7%)	48/97 (49.5%)
No Desire	21/67 (31.3%)	9/30 (30.0%)	30/97 (30.9%)
Lack of Resources	14/67 (20.9%)	5/30 (16.7%)	19/97 (19.6%)
Unsure Effectiveness	9/67 (13.4%)	5/30 (16.7%)	14/97 (14.4%)
Procrastination	9/67 (13.4%)	3/30 (10.0%)	12/97 (12.4%)
Distractions	7/67 (10.4%)	4/30 (13.3%)	11/97 (11.3%)
Hard to Change	4/67 (6.0%)	5/30 (16.7%)	9/97 (9.3%)
Nobody Else Available	2/67 (3.0%)	4/30 (13.3%)	6/97 (6.2%)
Mental Health	5/67 (7.5%)	0/30 (0.0%)	5/97 (5.2%)
Schedule Conflicts	2/67 (3.0%)	3/30 (10.0%)	5/97 (5.2%)
Lack of Flow	3/67 (4.5%)	1/30 (3.3%)	4/97 (4.1%)
Variety Requirement	2/67 (3.0%)	2/30 (6.7%)	4/97 (4.1%)
Distance	3/67 (4.5%)	0/30 (0.0%)	3/97 (3.1%)
Introversion	3/67 (4.5%)	0/30 (0.0%)	3/97 (3.1%)
Monetary Restrictions	1/67 (1.5%)	1/30 (3.3%)	2/97 (2.1%)

The most commonly cited barrier was lack of time, with many students stating that their packed schedules and competing priorities prevented them from trying more effortful strategies. Another frequently mentioned obstacle was lack of resources, which for many students specifically referred to the unavailability of tools such as retrieval practice tests, which they perceived as beneficial but difficult to access unless provided by instructors. Additionally, a lack of desire to adopt new strategies was widely reported across both groups, with some students feeling unmotivated or believing that changing their study habits was unnecessary.

Other reported barriers included procrastination, which many students attributed to laziness, and uncertainty about effectiveness, with some participants expressing doubt about whether new strategies would actually improve their learning outcomes—particularly their grades. Distractions were another concern, with some students noting that spacing their study sessions made it easier for them to lose focus between intervals, in particular, some students stated that they believed it disrupted their ability to stay in the flow. Lastly, difficulty in changing habits was highlighted as a barrier, as students found it challenging to leave the familiar but less effective learning approaches.

4. Discussion

We examined whether an intervention on the effectiveness of learning strategies, consisting of instruction and direct experience with effective learning practices, could lead to long-term shifts in students' understanding and use of these strategies. Our findings indicated that, compared to the business-as-usual sections of Introduction to Psychology,

the intervention led to meaningful improvements in students' understanding and beliefs about the effectiveness of various learning strategies. Indeed, a key finding of the study was that students in the intervention group not only maintained their increased understanding of effective learning strategies over time but also continued applying them, in contrast to the decline observed in the business-as-usual group. This stability suggests that students who engaged in the intervention were able to internalize and implement at least some of the effective strategies without continued external guidance.

The persistence in students' understanding and use of effective learning strategies likely stems from the integration of both instruction and experience-based practice. The intervention enhanced students' knowledge through explicit teaching of the principles underlying effective strategies as well as by ensuring that students utilized many of those effective practices to study the material throughout the semester (e.g., ensuring that they were forced to study material about effective learning practices via active retrieval-based means at timepoints distributed throughout the entirety of the semester, etc.). The intervention thus also helped reinforce that these strategies are effective by providing opportunities to experience them firsthand. This dual approach addressed limitations noted in previous research, which suggests that experience-based approaches may improve behaviors but struggle to enhance understanding without explicit guidance (Pan & Rivers, 2023), and that instruction alone may lead to modest shifts in beliefs but often fails to produce long-term behavioral change (Balch, 2001; Dembo & Seli, 2004; Foerst et al., 2017). This combined approach promoted more enduring changes in students' knowledge and application of effective learning strategies.

Another notable finding was that a similar pattern was observed for ineffective strategies as well. In the immediate post-intervention period, students improved their understanding of effective strategies and reduced their use of less effective ones. This suggests that the intervention successfully facilitated an initial shift in both knowledge and behavior regarding ineffective strategies. However, while students maintained their improved understanding over time, the changes in behavior did not persist. In the long-term follow-up, some students reverted to their previous habits, and the reduction in reliance on ineffective strategies diminished.

While it is encouraging that students maintained their use of effective strategies over time, the lack of a substantial increase—along with a return to some ineffective strategies—demonstrates the ongoing challenge of translating knowledge into behavioral change. This pattern reflects a well-documented persistent challenge in prior intervention studies: even as students' understanding of effective strategies grows, they often continue to rely on familiar but ineffective approaches (Ariel & Karpicke, 2018; Brown-Kramer, 2022; Gurung & Burns, 2019; Yüksel et al., 2024). To better understand these challenges, we examined students' self-reported experiences through open-ended questions. These questions provided additional information on how students conceptualize effective learning and the barriers they encounter when attempting to implement effective strategies.

One critical finding that may play a role in the observed mismatch between knowledge and behavior was with respect to the participants' stated goals in courses. While most experimental work starts from the premise that students seek to learn, this was not a major goal of our participants. Instead, many students stated that their strategies were effective because they helped them achieve good grades. Their primary goal was performing well on exams and assignments, whereas the intervention aimed to promote changes in learning strategies that support long-term retention (see Soderstrom & Bjork, 2015). This misalignment in goals may make students less likely to implement some of the effective strategies suggested by the intervention. For instance, rereading can be relatively effective for immediate recall (van den Broek et al., 2014), reinforcing the belief that it is a useful

strategy, even though it may not support long-term learning as well as strategies like spacing or retrieval practice.

This finding offers useful directions for the design of future studies and interventions, highlighting the importance of addressing the misalignment between students' goals and long-term learning outcomes. One way to support this shift might be to incorporate goal-setting components into the intervention, explicitly encouraging students to align their study strategies with long-term retention rather than short-term performance. At the same time, instructors play a crucial role in shaping students' learning priorities. Structuring courses to emphasize conceptual understanding, mastery, and sustained retention, rather than focusing solely on one-time exam performance, may encourage students to shift their goals toward deeper learning. Given that fully changing students' mindsets away from grade-oriented goals may be difficult, it could also be beneficial to frame effective strategies in terms of both learning benefits and potential grade improvements. For example, future interventions could make more explicit how using these strategies may lead to better performance on exams—an outcome many students value. Future research should examine alternative ways of creating curricula that prioritizes long-term mastery while still appealing to students' performance-related motivations.

Another valuable finding from the open-ended questions was that even among students who were more aligned with the intervention's goal of long-term retention, many struggled with implementing effective strategies in their daily study routines. A common challenge they faced was feeling limited in their available time. Students frequently mentioned that time constraints made it difficult to engage in more effortful but effective strategies, such as spaced learning. To address these challenges, some students mentioned using time management tools; however, others struggled with the additional effort required to plan their studying. Explicit instruction in time management could be a key component of future interventions aimed at promoting strategies like spacing and interleaving.

While addressing these barriers within the intervention is one way to produce further behavioral changes, some obstacles may also require fostering a broader learning environment that actively supports good practices across multiple classrooms. In particular, the low reported frequency of use of some learning strategies may not reflect a lack of understanding or motivation, but rather a lack of access to the necessary resources. For instance, because limited access—especially for retrieval practice—was frequently cited as a barrier, instructors could incorporate cumulative weekly quizzes into coursework to provide structured self-testing opportunities while also facilitating interleaving and spacing. Addressing these resource limitations outside the intervention itself may therefore be key to increasing the use of effective strategies. When students consistently engage with these practices in the context of their regular coursework, they are more likely to develop long-lasting habits that support sustained learning (see for review Zimmerman, 2002). In turn, the impact of interventions like ours may be further strengthened when reinforced across different learning environments.

Despite the effectiveness of the intervention and the valuable insights gained from this study, it is also important to acknowledge its limitations. Although we reached out to a large pool of students, the final sample size was small relative to the entire eligible population as well as relative to the previously published work. Moreover, while the participation rates for both the intervention and business as usual groups were within a few percentage points of each other, because the rate was slightly higher in the intervention group, additional caution with regard to drawing strong causal inferences regarding the impact of the intervention is warranted. The voluntary nature of participation may also have introduced selection bias, as those who enrolled could differ in meaningful ways from the broader student population. Although the data suggest that there is strong commonality

between at least the subset of participants who participated in the current work and in the Yüksel et al. (2024) study, these factors nonetheless limit the generalizability of our findings. Additionally, while the natural setting of this study allowed us to assess the intervention's real-world impact across five years, it also meant that we could not track participants as systematically as in a controlled experiment. Furthermore, given the relatively small sample size, we were cautious not to overinterpret differences at the level of individual learning strategies. Given the many nuances associated with individual strategies (e.g., where the degree of effectiveness may depend on implementation or contextual details), future research should prioritize methods that allow for deeper strategy-level analyses.

Another limitation of the current design is the potential for order effects, as the perceived effectiveness ratings were collected prior to reported frequency of use. This raises the possibility that participants' judgments of effectiveness may have influenced how frequently they reported using each strategy. While we did observe a number of cases where ratings of effectiveness and frequency of use diverged (e.g., cramming was rated as relatively ineffective, but also commonly used), it is possible that these are underestimates of the true degree of divergence. Because the sheer number of items makes full counterbalancing impossible, it would be valuable for future work to consider alternative approaches to reduce the possibility that order effects play a role in observed results (e.g., by counterbalancing effectiveness ratings and frequency of use ratings, without item-level counterbalancing). Finally, we measured only the self-reported frequency of strategy use; future studies should aim to capture actual behavioral changes. Future research might benefit from longitudinal designs that follow individual students over time to better capture changes in understanding, strategy use, and obstacles.

In conclusion, our findings demonstrate that while it is possible to foster sustainable changes in students' understanding and use of effective learning strategies, achieving long-term behavioral change requires addressing both cognitive and logistical barriers. Future interventions should educate students on effective strategies and provide concrete tools, support, and reinforcement to help them integrate these strategies into their routines. Furthermore, instructors play a crucial role in embedding these strategies into coursework in order to make their adoption easier and more enduring. Actively engaging students in the learning process and reinforcing these strategies over time can support their consistent use. By shifting the focus from short-term performance to long-term mastery, future efforts can bridge the gap between knowledge and behavior, ultimately fostering better self-regulated learning practices.

Author Contributions: Conceptualization, E.M.Y., C.S.G., and H.A.V.; methodology, all authors contributed; software, E.M.Y.; validation, all authors contributed; formal analysis, E.M.Y.; investigation, E.M.Y.; resources, C.S.G. and H.A.V.; data curation, E.M.Y.; writing—original draft preparation, E.M.Y.; writing—review and editing, C.S.G. and H.A.V.; visualization, E.M.Y.; supervision, C.S.G. and H.A.V.; project administration, E.M.Y.; funding acquisition, C.S.G. and H.A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of University of Wisconsin-Madison (protocol code 2020-1601 on 29 December 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original data presented in the study and the scripts for data analyses are openly available in https://osf.io/kcduz/?view_only=8ba58424b68743c9a39a6e7b66c34516 accessed on 24 April 2025.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Appendix A.1

Table A1. Perceived effectiveness ratings for each learning strategy in Study 1.

Learning Strategies	Intervention			Business-as-Usual			Group × Time *p*-Value
	Pre-Intervention	Post-Intervention	Long-Term	Pre-Intervention	Post-Intervention	Long-Term	
Reading Textbook	6.36	4.06	2.88	6.79	6.69	5.83	0.052
Taking Tests	8.36	9.88	9.71	9.15	9.53	9.28	0.046 *
Using Life Examples	7.82	9.12	8.29	8.06	8.62	7.91	0.547
Looking at Notes	7.91	6.59	5.29	8.41	7.49	7.91	0.030 *
Highlighting Notes	5.64	2.53	3.29	7.15	4.51	5.19	0.773
Creating Outlines	6.27	6.41	6.53	7.76	5.38	7.02	0.588
Copying Notes	6.00	4.94	5.53	6.53	6.09	6.91	0.526
Highlighting Textbook	5.73	1.71	2.12	6.47	4.20	4.63	0.150
Cramming	4.73	3.29	3.65	3.79	3.84	3.85	0.421
Summarizing Notes	8.00	7.00	8.35	8.38	7.76	7.85	0.283
Using Flashcards	6.82	7.12	7.24	7.59	7.76	7.59	0.718
Group Study	6.36	7.65	6.88	6.76	6.11	5.98	0.365

Each strategy was analyzed using a linear mixed-effects model with time (coded as pre-intervention, post-intervention, and long-term follow-up), section (intervention and business-as-usual) and the interaction term as fixed effects, and a random intercept for part. An asterisk (*) indicates that the *p*-value is statistically significant.

Table A2. Reported frequency of use for each learning strategy in Study 1.

Learning Strategies	Intervention			Business-as-Usual			Group × Time *p*-Value
	Pre-Intervention	Post-Intervention	Long-Term	Pre-Intervention	Post-Intervention	Long-Term	
Reading Textbook	5.91	3.59	3.41	6.24	6.24	4.74	0.59
Taking Tests	8.00	9.52	8.47	8.50	7.53	7.21	0.17
Using Life Examples	6.82	7.76	7.12	7.79	6.75	6.15	0.09
Looking at Notes	8.91	8.00	8.35	8.47	8.46	8.33	0.74
Highlighting Notes	6.00	2.47	4.00	7.15	4.77	4.37	0.47
Creating Outlines	5.90	4.00	5.35	6.64	4.48	5.04	0.42
Copying Notes	5.45	5.53	6.41	5.09	4.93	6.35	0.81
Highlighting Textbook	5.81	1.23	2.11	5.14	2.31	2.54	0.44
Cramming	6.00	6.35	6.17	5.44	7.72	6.84	0.36
Summarizing Notes	7.36	5.94	7.35	7.70	6.71	6.35	0.19
Using Flashcards	5.54	3.64	4.47	5.50	5.68	4.73	0.99
Group Study	5.82	3.53	5.29	5.18	3.93	4.13	0.57

For both the business-as-usual and intervention groups, assessment time points are labeled based on their timing relative to the intervention period. "Pre-intervention" refers to assessments conducted before the intervention occurred, "post-intervention" to those conducted immediately after, and "long-term" to those conducted approximately six months later. Each strategy was analyzed using a linear mixed-effects model with time (coded as pre-intervention, post-intervention, and long-term follow-up), section (intervention and business-as-usual) and the interaction term as fixed effects, and a random intercept for time.

Appendix A.2

Table A3. Responses to open-ended questions: obstacles in Study 1.

Response	Business-as-Usual	Intervention	Total
Not Enough Time	2	0	2
No Desire	1	0	1
Lack of Resources	4	1	5
Unsure Effectiveness	13	3	16
Procrastination	3	1	4
Distractions	11	5	16
Hard to Change	2	0	2
Nobody Else Available	1	1	2
Mental Health	1	2	3

Table A4. Responses to open-ended questions: current strategies in Study 1.

Response	Business-as-Usual	Intervention	Total
Retrieval Practice	8	3	11
Rereading	1	1	2
Cramming	5	0	5
Time Management	3	1	4
Spacing	1	0	1
Using Flashcards	3	1	4
Studying in Advance	9	3	12
Copying Notes	5	0	5
Creating Outline	2	1	3
Group Study	1	0	1
Active Notetaking	5	0	5
Metacognitive Monitoring	1	0	1
Watching Videos	3	0	3
Self-Explanation	13	5	18
Summarization	3	1	4
Using Planner	4	2	6
Visualization	3	0	3
Environment Optimization	1	0	1
Error Analysis	1	0	1
Attending Office Hours	5	0	5
Studying in Different Places	3	1	4
Highlighting	1	0	1
Learning From Examples	1	0	1
Listening Audiobooks	1	0	1
Interleaved Practice	1	0	1

Appendix B

Appendix B.1

Table A5. Number of participants per semester, broken down by group.

	Number of Participants		
Semester	**Business-as-Usual**	**Intervention**	**Total**
2019–2020 Fall	10	0	10
2019–2020 Spring	2	0	2
2020–2021 Fall	10	5	15

Table A5. *Cont.*

	Number of Participants		
Semester	**Business-as-Usual**	**Intervention**	**Total**
2020–2021 Spring	2	3	5
2021 Summer	2	0	2
2021–2022 Fall	15	4	19
2021–2022 Spring	9	0	9
2022 Summer	1	0	1
2022–2023 Fall	8	22	30
2022–2023 Spring	9	0	9

Appendix B.2

Table A6. Perceived effectiveness ratings for each learning strategy in Study 2.

Learning Strategies	Intervention	Business-as-Usual	*p* Value
Reading Textbook	3.74	5	0.87
Looking at Notes	6.41	7.44	0.94
Copying Notes	5.24	6.1	0.20
Summarizing Notes	6.44	7.12	0.80
Taking Tests	9.26	8.81	0.06
Highlighting Notes	3.09	4.53	0.23
Highlighting Textbook	2.79	3.66	0.73
Using Flashcards	7.09	6.87	0.24
Using Life Examples	8.29	7.5	0.10
Creating Outlines	7	6.53	0.098
Cramming	3.82	4.54	0.278
Group Study	7.32	5.6	0.01 *
Studying in Different Places	8.15	5.81	0.0001 **
Spacing	8.15	7.29	0.45
Self-Explanation	9.29	8.59	0.13
Interleaving	5.88	4.37	0.002 **

A double asterisk (**) indicates statistical significance at $p < 0.001$, and a single asterisk (*) indicates statistical significance at $p < 0.05$.

Table A7. Reported frequency of use for each learning strategy in Study 2.

Learning Strategies	Intervention	Business-as-Usual	*p* Value
Reading Textbook	4.24	4.43	0.17
Looking at Notes	8.15	8.35	0.86
Copying Notes	5.59	5.82	0.48
Summarizing Notes	5.47	5.99	0.69
Taking Tests	8.44	7.46	0.15
Highlighting Notes	3.06	5.04	0.21
Highlighting Textbook	2.5	3.18	0.71
Using Flashcards	5.12	5.29	0.76
Using Life Examples	6.62	6.32	0.73
Creating Outlines	5.56	5.49	0.57
Cramming	6.29	6.43	0.51
Group Study	5.62	4.35	0.11
Studying in Different Places	7.29	5.37	0.01 *
Spacing	6.44	5.84	0.56
Self-Explanation	6.88	5.79	0.14
Interleaving	6.41	4.79	0.007 **

A double asterisk (**) indicates statistical significance at $p < 0.001$, and a single asterisk (*) indicates statistical significance at $p < 0.05$.

Appendix C

Professor Haskins is a professor of Art History. For the next assignment, she has asked the students in her class to learn to recognize paintings by three different artists—Jan Blencowe, Richard Lindenberg, and Rae O'Shea. Professor Haskins has two online slideshows available to teach students to recognize these painters. Each slide show contains the same 6 paintings from each artist, but differ in the order that the paintings are presented.

In Slideshow #1, all six paintings from a given artist are presented together before moving to the next artist (i.e., Blencowe-Blencowe-Blencowe-Blencowe-Blencowe-Blencowe-Lindenberg-Lindenberg-Lindenberg-Lindenberg-Lindenberg-Lindenberg-O'Shea-O'Shea-O'Shea-O'Shea-O'Shea-O'Shea).

In Slideshow #2, the slideshow rotates between the three artists (i.e., Blencowe-Lindenberg-O'Shea- Blencowe-Lindenberg-O'Shea-Blencowe-Lindenberg-O'Shea-Blencowe-Lindenberg-O'Shea- Blencowe-Lindenberg-O'Shea-Blencowe- Lindenberg-O'Shea).

The students are allowed to choose for themselves which slideshow to use to study.

If you were a student in Professor Haskins' class, which slideshow would you choose?

Definitely Slideshow #1 2 3 4 5 6 7 8 9 Definitely Slideshow #2

Figure A1. Example scenario question (interleaving vs. blocking).

Survey Prompt: In the following section you will see a set of questions related to your study activities. For each, choose the answer that is most true for you. Remember, there are no right or wrong answers.

1. How frequently do you expect to use each of the following studying methods this semester? (circle one)

Not at all—Very frequently

Rereading the textbook 1 2 3 4 5 6 7 8 9 10

Looking over notes 1 2 3 4 5 6 7 8 9 10

Copying notes 1 2 3 4 5 6 7 8 9 10

This figure displays the original prompt, a sample survey question, and a partial list of learning strategies presented to participants. Only a subset of strategies is shown here for illustrative purposes.

Figure A2. Example Survey Prompt, Question, and Sample Learning Strategies.

References

Ahmed, S. F., Tang, S., Waters, N. E., & Davis-Kean, P. (2019). Executive function and academic achievement: Longitudinal relations from early childhood to adolescence. *Journal of Educational Psychology*, *111*(3), 446–458. [CrossRef]

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied*, *24*(1), 43–56. [CrossRef] [PubMed]

Balch, W. R. (2001). Study tips: How helpful do introductory psychology students find them? *Teaching of Psychology*, *28*(4), 272–274. [CrossRef]

Balch, W. R. (2006). Encouraging distributed study: A classroom experiment on the spacing effect. *Teaching of Psychology*, *33*(4), 249–252.

Bartoszewski, B. L., & Gurung, R. A. R. (2015). Comparing the relationship of learning techniques and exam score. *Scholarship of Teaching and Learning in Psychology*, *1*(3), 219–228. [CrossRef]

Bowen, R. W., & Wingo, J. M. (2012). Predicting success in introductory psychology from early testing: High consistency and low trajectory in multiple-choice test performance. *North American Journal of Psychology*, *14*(3), 419–434. [CrossRef]

Brown-Kramer, C. R. (2022). Teaching students to learn better in introductory psychology: A replication and extension study. *Teaching of Psychology*, *49*(2), 108–117. [CrossRef]

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4–5), 514–527. [CrossRef]

Cassidy, S. (2012). Exploring individual differences as determining factors in student academic achievement in higher education. *Studies in Higher Education*, *37*(7), 793–810. [CrossRef]

Cathey, C. L., Visio, M. E., Whisenhunt, B. L., Hudson, D. L., & Shoptaugh, C. F. (2016). Helping when they are listening: A midterm study skills intervention for introductory psychology. *Psychology Learning & Teaching*, *15*(3), 250–267. [CrossRef]

Chen, G., Gully, S. M., Whiteman, J. A., & Kilcullen, R. N. (2000). Examination of relationships among trait-like individual differences, state-like individual differences, and learning performance. *Journal of Applied Psychology*, *85*(6), 835–847. [CrossRef] [PubMed]

Dembo, M. H., & Seli, H. P. (2004). Students' resistance to change in learning strategies courses. *Journal of Developmental Education*, *27*(3), 2.

DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition*, *32*(6), 945–955. [CrossRef]

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. [CrossRef] [PubMed]

Einstein, G. O., Mullet, H. G., & Harrison, T. L. (2012). The testing effect: Illustrating a fundamental concept and changing study strategies. *Teaching of Psychology*, *39*(3), 190–193. [CrossRef]

Foerst, N. M., Klug, J., Jöstl, G., Spiel, C., & Schober, B. (2017). Knowledge vs. action: Discrepancies in university students' knowledge about and self-reported use of self-regulated learning strategies. *Frontiers in Psychology*, *8*, 1288. [CrossRef]

Fowler, R. L., & Barker, A. S. (1974). Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology*, *59*(3), 358–364. [CrossRef]

Gurung, R. A. R., & Burns, K. (2019). Putting evidence-based claims to the test: A multi-site classroom study of retrieval practice and spaced practice. *Applied Cognitive Psychology*, *33*(5), 732–743. [CrossRef]

Gurung, R. A. R., Weidert, J., & Jeske, A. (2010). Focusing on how students study. Journal of the Scholarship of Teaching and Learning 10*(1)*, 28–35. Available online: https://scholarworks.iu.edu/journals/index.php/josotl/article/view/1734 (accessed on 1 June 2021).

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*(1), 126–134. [CrossRef]

Hassanbeigi, A., Askari, J., Nakhjavani, M., Shirkhoda, S., Barzegar, K., Mozayyan, M. R., & Fallahzadeh, H. (2011). The relationship between study skills and academic performance of university students. *Procedia—Social and Behavioral Sciences*, *30*, 1416–1424. [CrossRef]

Hattie, J. A., & Donoghue, G. M. (2018). A model of learning: Optimizing the effectiveness of learning strategies. In *Contemporary theories of learning* (pp. 97–113). Routledge. [CrossRef]

Hoon, P. W. (1974). Efficacy of three common study methods. *Psychological Reports*, *35*(3), 1057–1058. [CrossRef]

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. [CrossRef] [PubMed]

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224. [CrossRef]

Lyons, K. E., & Zelazo, P. D. (2011). Monitoring, metacognition, and executive function: Elucidating the role of self-reflection in the development of self-regulation. *Advances in Child Development and Behavior*, *40*, 379–412. [CrossRef]

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462–476. [CrossRef]

Pan, S. C., Flores, S. R., Kaku, M. E., & Lai, W. H. E. (2025). Interleaved practice enhances grammar skill learning for similar and dissimilar tenses in Romance languages. *Learning and Instruction*, *95*, 102045. [CrossRef]

Pan, S. C., & Rivers, M. L. (2023). Metacognitive awareness of the pretesting effect improves with self-regulation support. *Memory & Cognition 51*, 1461–1480. [CrossRef]

Ponce, H. R., Mayer, R. E., & Méndez, E. E. (2022). Effects of learner-generated highlighting and instructor-provided highlighting on learning from text: A meta-analysis. *Educational Psychology Review*, *34*(2), 989–1024. [CrossRef]

R Core Team. (2017). *R: A language and environment for statistical computing*. Available online: https://www.R-project.org/ (accessed on 24 April 2025).

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. [CrossRef]

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. [CrossRef] [PubMed]

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199. [CrossRef]

Usdansky, M. L., McDaniel, M. A., Razza, R., Lopoo, L. M., Tillotson, J. W., & Granato, R. (2024). Coaching to learn: Motivating students to adopt and adapt effective learning strategies. Learning Assistance Review *(*TLAR*)*, *28*(3), 223–256.

van den Broek, G. S., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*(7), 803–812. [CrossRef]

Witherby, A. E., Babineau, A. L., & Tauber, S. K. (2025). Students', teachers', and parents' knowledge about and perceptions of learning strategies. *Behavioral Sciences*, *15*(2), 160. [CrossRef] [PubMed]

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*(7), 918. [CrossRef]

Yüksel, E. M., Green, C. S., & Vlach, H. A. (2024). Effect of instruction and experience on students' learning strategies. *Metacognition Learning 19*, 345–364. [CrossRef]

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, *41*(2), 64–70. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Point Values on Scoring Rubrics Influence Self-Regulated Learning for STEM Material

Morgan D. Shumaker [1,*], Michelle L. Rivers [1,2] and Sarah K. Tauber [1]

[1] Department of Psychology, Texas Christian University, Fort Worth, TX 76129, USA; mrivers@scu.edu (M.L.R.); uma.tauber@tcu.edu (S.K.T.)
[2] Department of Psychology, Santa Clara University, Santa Clara, CA 95053, USA
* Correspondence: m.shumaker@tcu.edu

Abstract

We examined whether point value information on a scoring rubric influences learners' study time and concept selection when learning educationally relevant STEM information. Participants (N = 92) engaged in the self-regulated study of five concepts in mineral formation—geological processes, inorganic substances, compounds, elements, and crystalline solids—while having access to a scoring rubric that contained varying point values (concepts were worth 12, 8, or 4 points), uniform point values (all concepts were worth 8 points), or no point values for each concept. Participants chose how long to study and how many times to study each of the concepts. Concepts were selected for study more times when they were high-value than low-value on a grading rubric, an effect that was stronger for some concepts relative to others. Concepts were also studied slightly longer when they were high-value compared to low-value on a rubric. Our findings are consistent with value-directed remembering and agenda-based regulation and suggest that learners can use scoring rubrics to guide their decisions during learning.

Keywords: self-regulated learning; value-directed remembering; rubrics; education

1. Introduction

When studying for a course, students must self-regulate their learning by making decisions about what material to study and for how long to study it (Kornell & Bjork, 2007). Sometimes, instructors give students materials to help them guide their self-regulated learning such as a study guide or scoring rubric, which can make the instructors' expectations for the assignment clear (Panadero et al., 2014). Students often aim to maximize their performance and, therefore, may use whatever resource their instructors gave them (e.g., a scoring rubric) to create an agenda by which to guide their self-regulated learning.

The agenda-based regulation framework posits that students assess task constraints prior to study and construct an agenda that allows them to achieve the task goal within those constraints (Ariel et al., 2009, 2015). Specifically, students use their agenda to select which items to study and to allocate study time to those items when self-regulating their learning (Ariel et al., 2015). Various factors can influence students' agendas; one of which is the reward structure of the task. That is, students can construct an agenda aimed at maximizing their potential reward (Ariel et al., 2009).

One type of reward structure occurs when some of the material is deemed more valuable than other material (Middlebrooks & Castel, 2018). For example, an instructor might

state that certain material will appear on a test, whereas other material (e.g., supplementary) will not. Prior work has found that when learning material that differs with respect to value, learners tend to spend more time studying (Soderstrom & McCabe, 2011) and choose to study (Middlebrooks & Castel, 2018) higher-value over lower-value information. Prioritizing high-value items over low-value items often leads to superior memory for the high-value compared to low-value items, an effect known as value-directed remembering (Knowlton & Castel, 2022; Castel et al., 2009, 2011, 2013).

If learners' aim to maximize their score on an assignment where some material is worth more points than other material, focusing on high-value content would allow them to optimize their performance and achieve the task goal (Murphy & Castel, 2021b). In one study, when test items varied in point value (1 point or 5 points), participants more frequently chose to study the 5-point items, even though these items were more difficult to learn and likely required more time (Ariel et al., 2009). This suggests that students use the task's reward structure to guide their study strategies, prioritizing higher-value items to self-regulate their learning.

Using point values to guide learning is consistent with both the agenda-based regulation framework and the value-directed remembering framework. Much research has investigated agenda-based regulation (Dunlosky et al., 2011) and value-directed remembering (for a review, see Knowlton & Castel, 2022). However, most research on value-directed remembering focuses on memory outcomes (e.g., Castel et al., 2009, 2011, 2013) and encoding mechanisms (Schwartz et al., 2023; Hennessee et al., 2019), and uses materials such as word lists (e.g., Middlebrooks & Castel, 2018; but see Murphy & Castel, 2021a; Friedman et al., 2015). Additionally, most research that examines agenda-based regulation uses materials such as word pairs (e.g., Ariel et al., 2009). Thus, little is known about how value influences learners' study choices when engaging with educationally relevant materials, such as those encountered in STEM courses like geology.

One common scenario in which learners may encounter value-associated information is on a scoring rubric. To illustrate, instructors might provide a rubric to help students prepare a final research paper, indicating that the introduction section is worth 30 out of a possible 100 points on the assignment, whereas the method section is worth only 20 points. In this way, some components of the assignment carry more weight than others. This pedagogical strategy is presumably intended to direct students' attention and effort toward concepts the instructor considers most important for developing content knowledge. From the agenda-based regulation perspective, a rubric should help shape learners' study agendas and inform their study choices. If students aim to earn the highest possible score on the final paper, using the scoring rubric to prioritize high-value sections (e.g., the introduction) over low-value sections (e.g., the method) would be a strategic and beneficial approach.

Scoring rubrics can help learners better understand instructor expectations and support the development of more strategic study plans (Panadero et al., 2014; Reddy & Andrade, 2010). Additionally, giving learners access to a scoring rubric can increase their self-reported self-regulated learning (Camargo Salamanca et al., 2024). Specifically, learners who use rubrics while preparing for assignments report greater engagement in metacognitive processes such as the monitoring, planning, and evaluation of their learning compared to those without rubric access (Panadero & Romero, 2014). Learners' self-reports also indicate that access to scoring rubrics enhances metacognitive engagement during study. Further, using rubrics as a self-assessment tool can foster better self-regulation, including increased use of learning strategies like memorization (Karaman, 2024). However, not all findings are consistent; some studies have found limited evidence that rubrics impact students' self-regulated study behaviors (Fraile et al., 2023).

One limitation of previous research on learners' rubric use is its heavy reliance on self-report measures, personal anecdotes (Panadero & Jonsson, 2020; Fraile et al., 2023), and varied experimental designs (Camargo Salamanca et al., 2024). Because self-reported study behavior can diverge from actual behavior, the degree to which scoring rubrics impact learners' actual study choices remains an open question. Additionally, prior work differs with respect to the type of rubric used. Some rubrics contain subjective criteria (e.g., "all or most of the instruction involves active engagement"), others convey value-based information (e.g., "active engagement is worth 4 points"; Andrade & Du, 2005), and some provide scaffolded instructions (e.g., "underline key phrases in the rubric with colored ink"; Andrade et al., 2010; Andrade, 2019). Thus, a systematic examination of how point values on a rubric influence self-regulated learning has yet to be conducted. Specifically, to our knowledge, no study has compared rubrics with varying point values (high, medium, low) to rubrics with uniform (i.e., equally weighted) point values or rubrics with no point values. These latter formats may occur in practice when (1) an instructor indicates that all parts of the assignment are equally important and thus equally weighted, or (2) the instructor indicates the material that needs to be included but does not convey their relative point values.

The current study examines how point values on scoring rubrics influence self-regulated learning with objective measures of learners' study choices when learning STEM material. Specifically, we examine how rubrics that differ in the point value information they provide affect two key behaviors: study time (the number of minutes learners spend reading a concept) and concept selection (the number of times a concept is chosen for study). Participants prepared for a writing assignment on mineral formation while viewing a rubric that indicated how many points each concept was worth on the assignment. Informed by the value-directed remembering framework, we hypothesized that high-value concepts would be chosen for study more times than lower-value concepts. We expected no difference in concept selection when concepts were associated with equal weight or no point values. Additionally, we expected that high-value concepts would be studied longer relative to lower-value concepts. Finally, we expected no differences in study time when concepts had no point values or if all concepts were weighted equally on the rubric.

2. Materials and Methods

This experiment was preregistered on the Open Science Framework (OSF) and can be accessed at https://osf.io/zua8s/?view_only=763e12df66184f8088121a83558477f8 (accessed on 27 March 2025).

2.1. Participants and Design

Sample size was determined by conducting an a priori power analysis using G*Power 3.1.9.7 (Faul et al., 2007) to detect a medium effect ($\eta^2 = 0.06$ or $f = 0.25$) with 95% power, an alpha error probability of 0.05, and a correlation among repeated measures of 0.20 in a mixed-model analysis of variance (repeated measures, within–between interaction) with four groups. The analysis revealed that we needed 136 participants (34 per group) to meet these criteria. Thus, data were collected from a total of 136 Texas Christian University undergraduate students who participated in the experiment for partial course credit. Participants' ages ranged from 15 to 28 years old ($M = 20.00$, $SD = 1.58$) and most identified as women (20 men, 2 non-binary). Guardian consent as well as participant consent was gathered for participants under 18 years of age. Participants were randomly assigned to one of four rubric groups. Two rubric groups—uniform points and no points—were provided with the concepts that would appear on the writing assignment, but concepts were not differentiated in terms of point value (see bottom panel of Figure 1). The other two rubric

groups—varied points and reverse-varied points—were provided with the concepts that would appear on the writing assignment and points values that varied between concepts (see top panel of Figure 1). The purpose of having two rubric groups was to counterbalance the concepts that were high-value and low-value. All participants studied the same five concepts about mineral formation (geological processes, inorganic substances, crystalline solids, elements, compounds). This research was approved by the Institutional Review Board at Texas Christian University.

Varied Points

Scoring Rubric	
Concept	**Point Value**
Geological Processes	4
Inorganic Substances	12
Crystalline Solids	12
Elements	4
Compounds	8
Total Points Possible	**40**

Reverse-varied Points

Scoring Rubric	
Concept	**Point Value**
Geological Processes	12
Inorganic Substances	4
Crystalline Solids	4
Elements	12
Compounds	8
Total Points Possible	**40**

Uniform Points

Scoring Rubric	
Concept	**Point Value**
Geological Processes	8
Inorganic Substances	8
Crystalline Solids	8
Elements	8
Compounds	8
Total Points Possible	**40**

No Points

Scoring Rubric
Concept
Geological Processes
Inorganic Substances
Crystalline Solids
Elements
Compounds

Figure 1. Rubric for each group.

2.2. Materials

Five STEM concepts in geology (geological processes, inorganic substances, crystalline solids, elements, and compounds) and the definitions for each were selected from the work of Ariel et al. (2021). Concept definitions were short paragraphs (~120 words) about how the concept relates to the process of mineral formation. Below is an example of the paragraph for geological processes:

> "A mineral is a crystalline solid formed through geological processes. A geologic process is a natural process that occurs in or on Earth and that shapes Earth's features, including volcanic activity; the movement of tectonic plates; the formation of sedimentary layers of sand and mud; and the folding of those layers until, deep in the Earth, they are exposed to high pressures and temperatures."

We created four rubrics that differed by point value. In the groups with varied points and reverse-varied points, two concepts were assigned a high value (12 points), two concepts were assigned a low value (4 points), and one concept was assigned a medium value (8 points). In the group with uniform points, each of the five concepts was assigned a medium value (8 points). Finally, for the group with no points, each concept was listed without a point value (see Figure 1). Concepts were displayed on each rubric in a fixed order from top to bottom.

2.3. Procedure

This experiment was conducted in a lab, in groups ranging from one to eight participants who were tested simultaneously. As a measure of prior knowledge, participants were first asked to type everything they knew about each of the five concepts into a response box. Next, participants were instructed to learn about the process of mineral formation by reading about the individual concepts of mineral formation in preparation for a writing

assignment similar to the one they had just completed (i.e., the prior knowledge test). Participants were also instructed to study as if they were preparing for an assignment in one of their classes and were informed that they would receive a rubric that includes each concept that would be on the final writing assignment. Then, participants were given an example rubric and informed that while studying, they could click on any of the five concepts, in any desired order, to read about them. Participants were informed that they had to study at least one concept but afterwards, they could either choose to move onto the writing assignment or continue studying the concepts as many times as desired.

Following the instructions, a rubric appeared at the top of the study screen with each concept listed as a button below in the same fixed order as on the rubric. To study a concept, participants clicked on the respective button to display a paragraph about how the concept related to mineral formation. Participants could spend as much time as desired reading about each concept. After studying a concept, participants clicked an arrow to move back to the study screen. On the study screen, participants were then given the option to continue studying any concept, in any order, for as long as desired or, to continue to the writing assignment.

Once participants were finished studying, they completed a writing assignment for which they were given the name of each of the five concepts of mineral formation and asked to type as much information as they could remember about each into a response box (see Figure 2). Participants wrote their answers in any order desired and took as much time as needed to complete the writing assignment; no feedback was provided.

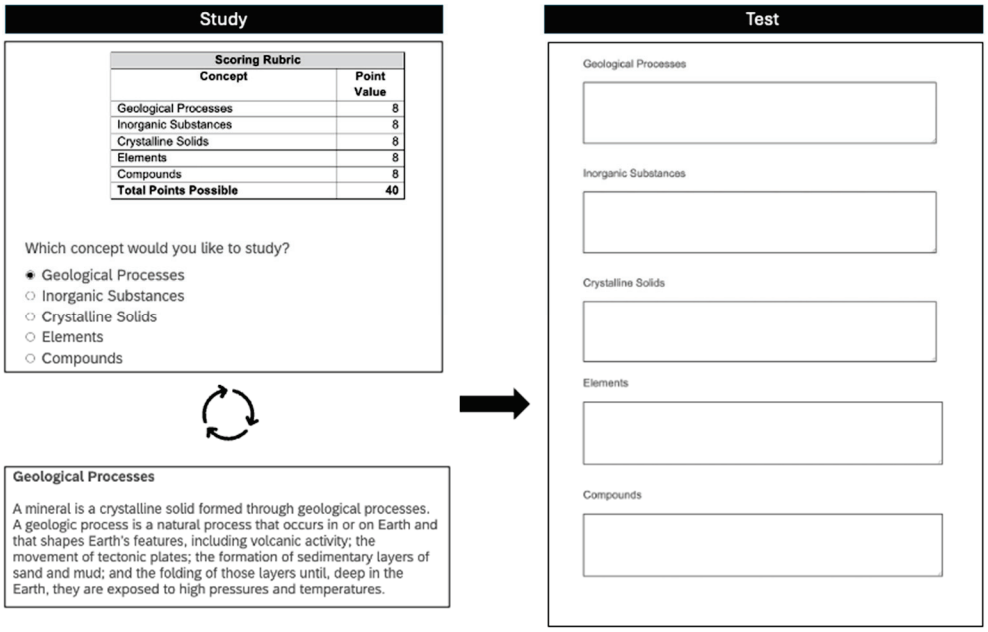


Figure 2. Task schematic. Participants began with the study screen, where they selected a concept to study (top left panel). They then read about the concept (bottom left panel) and returned to the study screen (top left panel) to make additional study selections or terminate study. Once participants terminated study, they completed the writing assignment (right panel).

Following the writing assignment, participants responded to the questions, "how did you determine that you were ready to complete the writing assignment?" and "did you use any strategies while preparing for the writing assignment?" by typing into a response box. Then, participants responded to the question, "how often do you get rubrics for writing assignments in your classes" and answered by selecting a multiple-choice option ranging from never to always. Participants also responded to the question "what is your opinion on having a rubric" and answered by selecting a multiple-choice option indicating that

they like receiving a rubric, do not like receiving a rubric, or do not have a preference on receiving a rubric. Finally, participants provided demographic information and were debriefed about the goal of the study.

2.4. Data Scoring

Each geological definition was composed of multiple key idea units. For example, the concept geological processes contained three idea units, so participants could earn up to three points. One idea unit was "natural process that occurs in or on Earth" and students earned 0.5 points for "natural processes" and 0.5 points for "that occurs in or on Earth." The second idea unit was "shapes Earth's features" for which students could earn 1 point by stating that and could earn 0.5 points (half credit) for stating something like "shapes the planet." For the third idea unit, participants could include "volcanic activity", "movement of tectonic plates", "formation of sedimentary layers of sand and mud", or "folding of those layers until, deep in the Earth, when they are exposed to high pressures and temperatures" and could earn 1 point for listing any of those terms or phrases (i.e., participants did not need to list each term or phrase). The number of idea units ranged from three to five per concept definition (see OSF for the scoring rubric: https://osf.io/zua8s/?view_only=763e12df66184f8088121a83558477f8; accessed on 27 March 2025). Participants were not penalized for minor spelling errors (e.g., quarts instead of quartz). Two independent raters, masked to group assignment, scored all responses. For each concept, the raters were trained to assign no credit (0 points), partial credit (0.5), or full credit (1 point) to each idea unit based on the correctness and completeness of the response. Agreement between raters was high, $r = 0.85$, $p < 0.001$, so scores were averaged between the two raters.

3. Results

Data and analyses are available on the Open Science Framework at https://osf.io/zua8s/?view_only=763e12df66184f8088121a83558477f8 (accessed on 27 March 2025). Data were analyzed in *R* (version 4.1.2; R Core Team, 2022) using the lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) packages. Bonferroni correction was applied to all follow-up analyses. We present the performance on concept selection, study time, short-answer test performance, and participants' responses to post-task questions. Data are organized so that concepts can be compared between value conditions (high value, low value, uniform points, no points). Because the concept of compounds was constantly worth medium points (8 points), it could not be included in the analyses to evaluate the impact of different point values (i.e., high vs. low) on any outcome measure. We examine the effect of rubric value and concept on concept selection, study time, and short-answer performance.

3.1. Concept Selection

We conducted a linear mixed effects model using rubric value and concept as interacting fixed effects and participant as a random effect to examine concept selection, defined as the number of times participants chose to study each concept. There was a significant main effect of rubric value on concept selection, $F(3,169) = 8.07$, $p < 0.001$, $\eta_p{}^2 = 0.96$, which is represented in the Average columns in Figure 3. Follow-up analyses revealed that concepts were selected for study significantly more often when associated with a high value ($M = 1.57$, $SE = 0.07$) compared with a low value on the rubric ($M = 1.30$, $SE = 0.07$), $p < 0.001$. In addition, concepts were selected for study more often when associated with the uniform points rubric ($M = 1.69$, $SE = 0.09$) compared with the low value on a rubric, $p = 0.007$. There was no effect of concept on concept selection, $F(3,396) = 0.50$, $p = 0.683$, and no interaction between value and concept, $F(9,3577) = 0.84$, $p = 0.582$.

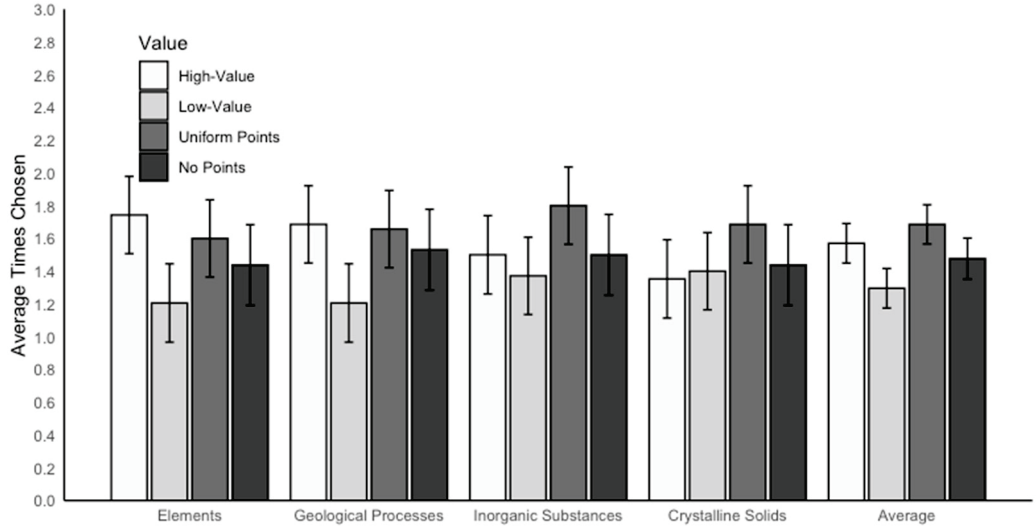


Figure 3. The average number of times each concept was chosen for study by rubric value. The concept of compounds is not included in this figure or associated analyses because it was used as a moderate-value control—i.e., it was not associated with either high or low values on any rubric.

Even though the interaction was not significant, we explored the effect of value on each concept separately to investigate the degree to which value impacted each. Linear regression models were conducted to examine the effect of rubric value on concept selection for each concept separately (see Figure 3). For elements, there was a main effect of rubric value on concept selection, $F(3,132) = 3.74$, $p = 0.013$, $\eta_p^2 = 0.08$. Follow-up analyses revealed that elements were selected for restudy significantly more often when associated with a high value ($M = 1.74$, $SE = 0.12$) relative to a low value on a rubric ($M = 1.21$, $SE = 0.12$), $p = 0.011$, $d = 0.77$. There were no other significant differences for the concept elements, ps ≥ 0.124. For geological processes, there was also a significant effect of rubric value on concept selection, $F(3,132) = 3.48$, $p = 0.01$, $\eta_p^2 = 0.07$. Follow-up analyses revealed that geological processes were selected for restudy significantly more often when associated with high value ($M = 1.69$, $SE = 0.12$) and uniform points ($M = 1.66$, $SE = 0.12$) relative to a low value on a rubric ($M = 1.21$, $SE = 0.12$), respectively, ps ≤ 0.045. There were no other significant differences for the concept geological processes, ps ≥ 0.348. There was no effect of rubric value on concept selection for inorganic substances, $F(3,132) = 2.23$, $p = 0.087$, or for crystalline solids, $F(3,132) = 1.45$, $p = 0.230$.

3.2. Study Time

Study time was recorded in Qualtrics as the total number of seconds participants spent on the study screen displaying information about each concept (see Figure 2) and was later converted to minutes. We conducted a linear mixed effects model using rubric value and concept as interacting fixed effects and participant as a random effect to examine study time. Outliers were considered those with study times more than two *SD* above the mean on any concept; there were 18 outliers. Excluding these outliers yielded similar patterns of results; thus, the full sample was maintained. There was a significant main effect of rubric value on study time, $F(3,169) = 3.12$, $p = 0.028$, $\eta_p^2 = 0.90$, which is represented in the Average columns in Figure 4. Post hoc comparisons revealed that high-value concepts were studied marginally longer ($M = 0.88$, $SE = 0.08$) than low-value concepts ($M = 0.74$, $SE = 0.08$), $p = 0.059$; however, this difference did not reach statistical significance. There were no other significant differences between rubric value groups on study time, ps ≥ 0.276. There was also a significant effect of concept on study time, $F(3,396) = 10.19$, $p < 0.001$, $\eta_p^2 = 0.97$. Participants studied crystalline solids ($M = 1.03$, $SE = 0.06$) for longer than

geological processes (M = 0.76, SE = 0.06), inorganic substances (M = 0.82, SE = 0.06), and elements (M = 0.83, SE = 0.06), ps $\leq$ 0.004. There were no other significant differences between concepts on study time, ps > 0.999.

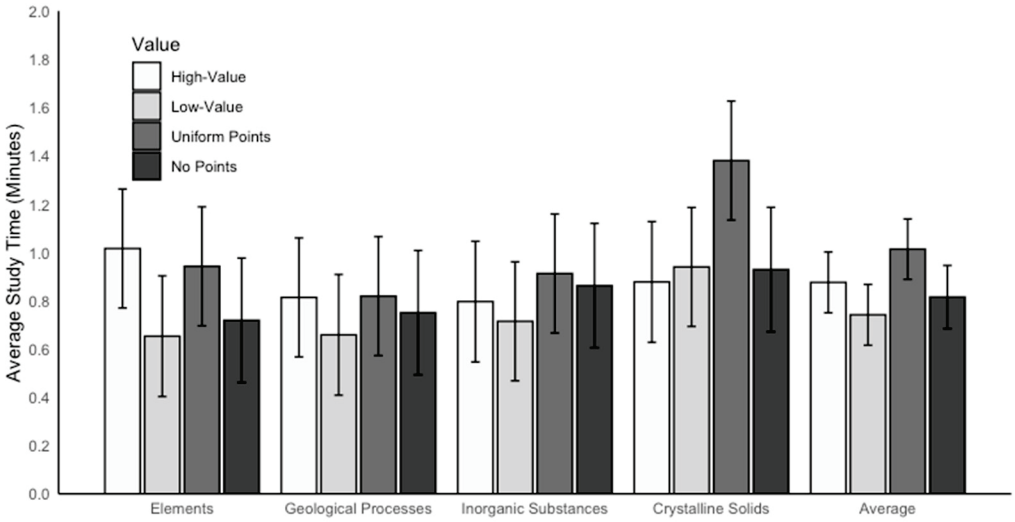


Figure 4. The average amount of time (in minutes) that each concept was studied by each rubric value. The concept of compounds is not included in this figure or associated analyses because it was used as a moderate-value control—i.e., it was not associated with either high or low values on any rubric.

Importantly, these effects were qualified by a significant interaction between rubric value and concept, F(9,338) = 2.20, p = 0.021.

The interaction was driven by study time differences for one concept—crystalline solids. Indeed, follow-up linear regression models revealed a marginal effect of rubric value on study time for crystalline solids, such that crystalline solids were studied longer when associated with uniform points (M = 1.38, SE = 0.15) compared with a low value on a rubric (M = 0.94, SE = 0.15), F(3,132) = 2.54, p = 0.058 (see Figure 4); this trend was not significant. There was no effect of rubric value on study time for elements: F(3,132) = 1.67, p = 0.176. There was also no effect of rubric value on study time for geological processes, F(3,132) = 0.44, p = 0.722, or for inorganic substances, F(3,132) = 0.65, p = 0.586.

3.3. Short-Answer Test Performance

We conducted a linear mixed effects model using rubric value and concept as interacting fixed effects and participant as a random effect to examine short-answer performance, defined as the correct proportion on the short-answer test. There was no effect of rubric value on short-answer performance, F(3,169) = 0.29, p = 0.832 (see the Average columns in Figure 5). However, there was a significant main effect of concept, F(3,396) = 23.07, p < 0.001, η_p^2 = 0.99 (see the Concept columns in Figure 5). Post hoc comparisons revealed that participants performed better on geological processes (M = 0.45, SE = 0.02) and inorganic substances (M = 0.44, SE = 0.02) than on crystalline solids (M = 0.30, SE = 0.02), ps < 0.001, ds $\geq$ 0.68. Additionally, participants performed better on geological processes and inorganic substances than on elements (M = 0.30, SE = 0.02), ps < 0.001, ds $\geq$ 0.67. There were no other significant differences between concepts on short-answer performance, ps > 0.999.

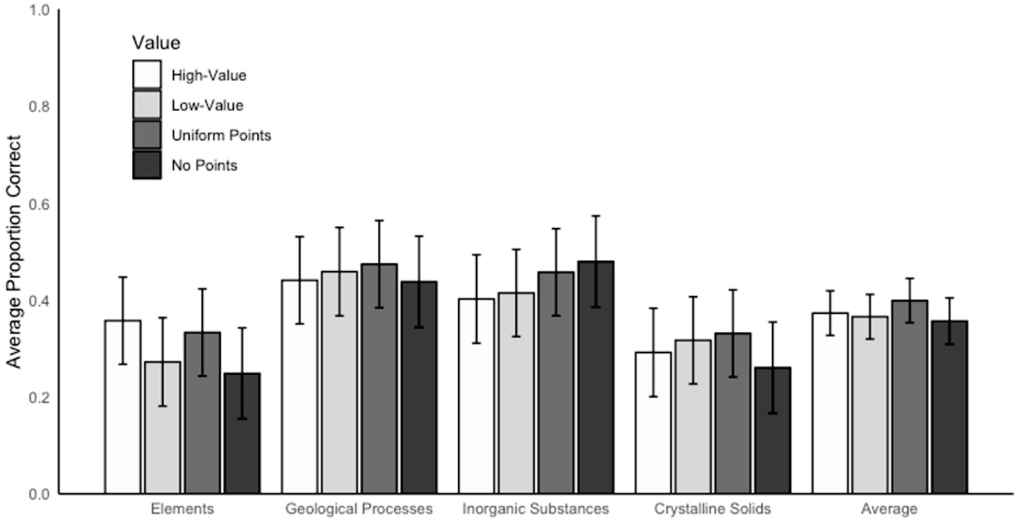


Figure 5. The average proportion correct on the short-answer test for each concept by rubric value. The concept of compounds is not included in this figure or associated analyses because it was used as a moderate-value control—i.e., it was not associated with either high or low values on any rubric.

We also conducted an exploratory moderation analysis to examine whether the relationship between rubric value and short-answer performance was moderated by concept selection. The overall model accounted for a significant portion of the variance in short-answer performance, $F(7,536) = 5.56$, $p < 0.001$, $R^2 = 0.07$. More frequent concept selection was associated with higher short-answer performance when no points were assigned to concepts, $b = 0.08$ ($SE = 0.03$), $t = 2.52$, $p = 0.012$. This relationship was not present when concepts were associated with high, low, or uniform point values, ps ≥ 0.725. Overall, concept selection did not moderate the effect of rubric value on short-answer performance, ps ≥ 0.400.

Finally, we conducted an exploratory moderation analysis to examine whether the relationship between rubric value and short-answer performance was moderated by study time. The overall model accounted for a significant portion of the variance in short-answer performance, $F(7,536) = 14.74$, $p < 0.001$, $R^2 = 0.16$. When no points were assigned to concepts, increased study time was associated with better short-answer performance, $b = 0.10$ ($SE = 0.03$), $t = 3.19$, $p = 0.002$. This relationship was not significant when concepts were assigned uniform points, $b = -0.04$ ($SE = 0.04$), $t = 0.99$, $p = 0.320$. However, study time was a significantly stronger predictor of short-answer performance when concepts were assigned high values, $b = 0.12$ ($SE = 0.04$), $t = 2.69$, $p = 0.007$, and when concepts were assigned low values, $b = 0.13$ ($SE = 0.05$), $t = 2.65$, $p = 0.008$, than when concepts were assigned no points. Thus, study time significantly moderated the relationship between rubric value and short-answer performance.

3.4. Responses to Post-Task Questions

Responses to "how did you determine that you were ready to complete the writing assignment" and "did you use any strategies while preparing for the writing assignment" were coded according to a scoring rubric. Many participants (47%) indicated that they determined they were ready to complete the writing assignment when they felt their recall was sufficient. Fewer participants reported that they decided they were ready for the test after they had read all the materials at least once (25%), when they felt confident about their knowledge (21%), when they could not remember any more information (4%), other (2%), or that they never felt ready (2%). There were no differences between rubric groups on how

participants determined they were ready for the final writing assignment, $\chi^2(15) = 9.66$, $p = 0.840$.

Additionally, many participants (51%) reported that they used multiple strategies (e.g., retrieval practice and key words) while preparing for the writing assignment. Fewer participants reported that they did not use any particular strategy (10%), or that they used one strategy. Specifically, some participants reported that they used repetition (10%), memorization (9%), key word or main idea identification (4%), connections to other material (4%), retrieval practice (3%), what they did not know (3%), mnemonics (2%), exclusively the point values on the rubric (1%), or another strategy such as chunking (4%). There were no differences between rubric groups on their self-reported strategy use, $\chi^2(30) = 20.10$, $p = 0.915$.

There was inconsistency in participants' reports of how often they receive rubrics in their classes. Some participants indicated that they receive rubrics for writing assignments in their actual classes (18% always receive rubrics; 24% sometimes receive rubrics), whereas other participants indicated they do not (8% do not often receive rubrics; 3% never receive rubrics). There were no differences between rubric groups on the frequency of receiving a rubric, $\chi^2(12) = 18.00$, $p = 0.114$. Most participants indicated that they like having a rubric (91%), and very few participants reported no preference on receiving a rubric (7%) or that they do not like receiving a rubric (2%). There were no differences between rubric groups on preference for receiving a rubric, $\chi^2(6) = 6.50$, $p = 0.369$.

4. Discussion

We examined the effect of scoring rubrics on learners' concept selection and study time when learning educationally relevant STEM material. Learners studied various concepts about mineral formation while having access to a rubric that indicated each concept was worth a different number of points or that all concepts were worth an equal number of points or a rubric that did not convey any information about point value for the concepts. Participants whose rubrics indicated that elements and geological processes were high-value chose to study those concepts more times than those whose rubrics indicated those concepts were low-value. Additionally, when studying geological processes, those in the uniform points (medium-value) group studied more times than those in the low-value group. There was also a trend such that when concepts were deemed high-value, participants studied numerically longer than when concepts were deemed low-value. There were no differences in concept selection or study time between those whose rubrics indicated that each concept was equally weighted or whose rubrics did not contain value. Thus, there is some evidence that learners made strategic choices (albeit imperfectly) based on the concept value when point values differed between concepts (high- vs. low-value). These outcomes are consistent with the value-directed remembering framework (Knowlton & Castel, 2022; Castel et al., 2009, 2011, 2013).

These findings are also in line with the agenda-based regulation framework (Ariel et al., 2009), suggesting that when concepts differed by point value on the rubric, learners used that information, at least to some extent, to create their agendas to meet their goal of maximizing their scores on the assignment. That is, when information on the rubric was useful (i.e., when concepts differed by point value), learners used the point values on the rubric to guide their concept selection (although, interestingly, many of them did not report doing so). This is, to our knowledge, the first objective evidence that learners do indeed use scoring rubrics as a basis for their study behavior when anticipating a writing assignment. Moreover, most learners report that they like having a rubric when completing an assignment. Future research could examine which types of rubrics best support students' self-regulated learning, as the rubrics in this experiment differ from rubrics used in prior

work (e.g., Andrade, 2019). The diversity in educators' approaches to rubric construction may contribute to interpretation challenges for learners and may directly impact the degree to which learners effectively use rubrics to direct their learning behavior.

Test performance was not influenced by value. This finding is inconsistent with prior work in value-directed remembering that has shown better memory for higher-value than lower-value material (e.g., Knowlton & Castel, 2022). This outcome is important for contextualizing the educational implications of the reported experiment because this suggests that even though value from the rubric influenced some of learners' study decisions, it did not have a meaningful impact on test performance. However, there are multiple reasons why we did not observe test performance effects. One reason may be that in typical value-directed remembering paradigms, learners need task experience (e.g., multiple trials) to learn to prioritize high-value over low-value information (e.g., Castel et al., 2009, 2011, 2013). Participants in our study only had one instance of having a scoring rubric while completing an assignment, so it will be important for future work to examine whether learners need practice using rubrics to use them more effectively. A second reason is that even though learners chose to study high-value concepts more times than other concepts, they may have needed to revisit concepts more times, spend more time with them, and use more effective learning strategies when learning the high-value concepts to observe test performance effects. A critical direction for the future will be to build on the reported outcomes to determine how to increase the quality of learners' study during self-regulated learning. One approach would be to conduct a classroom study in which students are more motivated to learn content that is directly needed for their class. In this context, value effects from a provided rubric may appear on test performance for the learners' class. At present, we recommend that educators provide rubrics that differentiate concepts by value so students can direct their study in preparation for class assignments and exams.

Additionally, the current study did not assess delayed test performance, which would be important both from the perspective of value-directed remembering effects and practical applications of the current research. Future work should examine how value portrayed on a rubric influences lasting learning. Finally, concept selection did not moderate the relationship between value and short-answer performance. However, increased study time was associated with better short-answer performance when concepts were associated with no points, and this relationship was even stronger when concepts were associated with high value sand low values. This suggests that students may use study time more strategically when rubrics emphasize certain concepts (e.g., high-value or low-value) over others (e.g., uniform points).

4.1. Limitations

This study is not without its limitations. First, the finding that learners selected to study concepts more times when they were high-value compared to low-value was driven primarily by some concepts rather than all concepts. This finding indicates that considering item characteristics will be critical for determining when and why learners' study decisions are influenced by value. Future research is needed to systematically evaluate the characteristics of complicated concepts (e.g., interest, applicability, abstractness) like the mineral formation concepts used in the reported experiment to determine what factors influence self-regulated learning when students have value information provided on a rubric. For instance, learners' study time may have been influenced by item difficulty, as indicated by differences in short-answer test performance. Specifically, crystalline solids was one of the more challenging concepts with lower test performance relative to geological processes and inorganic substances (ps ≤ 0.001).

Another limitation is that learners' study decisions may have also been influenced by habitual responding (e.g., studying from top to bottom; Dunlosky & Ariel, 2011). That is, concepts appeared in a fixed order on each rubric and on the concept selection screen for all participants. We adopted this method because it reflects common instructional practices. However, we suggest that follow-up research randomize the order of concepts on rubrics. This is important given that we found value effects to be stronger on learners' study choices for some concepts relative to others. Given that participants performed best on the first two listed concepts (geological processes and inorganic substances), it is possible that they better remembered these concepts simply because they studied them first. Indeed, exploratory analyses found that 67.65% of participants studied geological processes (the first listed concept) first. Further research is needed to disentangle the effects of habitual responding, learners' preferences for content, and value on short-answer performance.

4.2. Conclusions

Most importantly, we found evidence that learners can use rubrics when self-regulating their learning of complicated STEM concepts. Scoring rubrics that provide learners with information about value differences between concepts can influence learners' studying behavior in a laboratory setting. Future research could expand on these findings by examining whether scoring rubrics that differ with respect to value influence learners' studying behavior in a classroom setting, as students may be more motivated to use the scoring rubric to guide their learning. The current study also provides some evidence for value-directed and agenda-based regulation effects using educationally relevant material such as a scoring rubric. Scoring rubrics that contain information regarding how many points each topic is worth on an assignment may be beneficial for students' self-regulated learning. These findings have practical implications, as instructors could provide scoring rubrics where the material varies in point value to help students self-regulate their learning by guiding them toward prioritizing high-value over low-value information.

Author Contributions: Conceptualization, M.D.S., M.L.R. and S.K.T.; methodology, M.D.S.; formal analysis, M.D.S.; data curation, M.D.S.; writing—original draft preparation, M.D.S.; writing—review and editing, M.D.S., M.L.R. and S.K.T.; visualization, M.D.S.; supervision, S.K.T.; project administration, M.D.S.; funding acquisition, S.K.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition, Collaborative Grant number 220020483.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Texas Christian University (protocol code 1710-032-1809 and date of approval 31 July 2024).

Informed Consent Statement: Informed consent was obtained from the participants involved in this study.

Data Availability Statement: All materials, raw data, and analysis code are available from the Open Science Framework and can be freely accessed at https://osf.io/zua8s/?view_only=763e12df66184f8088121a83558477f8 (accessed on 27 March 2025).

Acknowledgments: We would like to thank Brooke Marvel, Chiara Thommarson, Maria Mendoza, Morgan Amberg, Niteeka Narayan, and Samantha Negrete for their assistance with data collection and Maria Mendoza and Morgan Amberg for their assistance with scoring.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Andrade, H. L. (2019). A critical review of research on student self-assessment. In *Frontiers in education* (Vol. 4, p. 87). Frontiers Media SA.

Andrade, H. L., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research, and Evaluation*, *10*, 3.

Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 199–214.

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General*, *138*(3), 432–447. [CrossRef] [PubMed]

Ariel, R., Karpicke, J. D., Witherby, A. E., & Tauber, S. K. (2021). Do judgments of learning directly enhance learning of educational materials? *Educational Psychology Review*, *33*, 693–712. [CrossRef]

Ariel, R., Price, J., & Hertzog, C. (2015). Age-related associative memory deficits in value-based remembering: The contribution of agenda-based regulation and strategy use. *Psychology and Aging*, *30*(4), 795–808. [CrossRef]

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. [CrossRef]

Camargo Salamanca, S. L., Parra-Martínez, A., Chang, A., Maeda, Y., & Traynor, A. (2024). The effect of scoring rubrics use on self-efficacy and self-regulation. *Educational Psychology Review*, *36*, 70. [CrossRef]

Castel, A. D., Balota, D. A., & McCabe, D. P. (2009). Memory efficiency and the strategic control of attention at encoding: Impairments of value-directed remembering in Alzheimer's disease. *Neuropsychology*, *23*(3), 297–306. [CrossRef]

Castel, A. D., Humphreys, K. L., Lee, S. S., Galván, A., Balota, D. A., & McCabe, D. P. (2011). The development of memory efficiency and value-directed remembering across the life span: A cross-sectional study of memory and selectivity. *Developmental Psychology*, *47*(6), 1553–1564. [CrossRef]

Castel, A. D., Murayama, K., Friedman, M. C., McGillivray, S., & Link, I. (2013). Selecting valuable information to remember: Age-related differences and similarities in self-regulated learning. *Psychology and Aging*, *28*(1), 232–242. [CrossRef]

Dunlosky, J., & Ariel, R. (2011). The influence of agenda-based and habitual processes on item selection during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 899–912. [CrossRef] [PubMed]

Dunlosky, J., Ariel, R., & Thiede, K. W. (2011). Agenda-based regulation of study-time allocation. In *Constructions of remembering and metacognition: Essays in honour of bruce whittlesea* (pp. 182–198). Palgrave Macmillan UK.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. [CrossRef] [PubMed]

Fraile, J., Gil-Izquierdo, M., & Medina-Moral, E. (2023). The impact of rubrics and scripts on self-regulation, self-efficacy and performance in collaborative problem-solving tasks. *Assessment & Evaluation in Higher Education*, *48*(8), 1223–1239.

Friedman, M. C., McGillivray, S., Murayama, K., & Castel, A. D. (2015). Memory for medication side effects in younger and older adults: The role of subjective and objective importance. *Memory & Cognition*, *43*, 206–215.

Hennessee, J. P., Patterson, T. K., Castel, A. D., & Knowlton, B. J. (2019). Forget me not: Encoding processes in value-directed remembering. *Journal of Memory and Language*, *106*, 29–39. [CrossRef]

Karaman, P. (2024). Effects of using rubrics in self-assessment with instructor feedback on pre-service teachers' academic performance, self-regulated learning and perceptions of self-assessment. *European Journal of Psychology of Education*, *39*(3), 2551–2574. [CrossRef]

Knowlton, B. J., & Castel, A. D. (2022). Memory and reward-based learning: A value-directed remembering perspective. *Annual Review of Psychology*, *73*(1), 25–52. [CrossRef]

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. [CrossRef]

Middlebrooks, C. D., & Castel, A. D. (2018). Self-regulated learning of important information under sequential and simultaneous encoding conditions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(5), 779–792. [CrossRef]

Murphy, D. H., & Castel, A. D. (2021a). Responsible remembering and forgetting as contributors to memory for important information. *Memory & Cognition*, *49*(5), 895–911.

Murphy, D. H., & Castel, A. D. (2021b). The role of attention and ageing in the retrieval dynamics of value-directed remembering. *Quarterly Journal of Experimental Psychology*, *75*(5), 954–968. [CrossRef]

Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (2014). Rubrics vs. self-assessment scripts: Effects on first year university students' self-regulation and performance. *Journal for the Study of Education and Development*, *37*(1), 149–183.

Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review*, *30*, 100329. [CrossRef]

Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education: Principles, Policy & Practice*, *21*(2), 133–148.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 15 February 2025).

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435–448.

Schwartz, S. T., Siegel, A. L., Eich, T. S., & Castel, A. D. (2023). Value-directed memory selectivity relies on goal-directed knowledge of value structure prior to encoding in young and older adults. *Psychology and Aging*, *38*(1), 30–48. [CrossRef]

Soderstrom, N. C., & McCabe, D. P. (2011). The interplay between value and relatedness as bases for metacognitive monitoring and control: Evidence for agenda-based monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1236–1242. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Review

Relational Reasoning and Educational Applications

Lindsey Engle Richland [1,*] and Hongyang Zhao [2]

[1] School of Education, University of California, Irvine, Irvine, CA 92697, USA
[2] Department of Educational Psychology, College of Education, University of Illinois Urbana-Champaign, Champaign, IL 61820, USA; zhaohy@illinois.edu
* Correspondence: l.richland@uci.edu

Abstract

As the global economy and sociopolitical, technological, climate, and international trade contexts continue to shift rapidly, educational goals and aims for skilled employment are changing accordingly. Research on the cognition of relational reasoning provides insight into training youth to handle these types of complex systems and multi-faceted problems. Relational reasoning is the ability to make inferences from the relationships within and between mental representations, and grounds many higher order cognitive actions such as generating solutions to novel problems, transferring insights from one context to another, and making creative leaps of inference that are central to these societal aims. This article synthesizes the theoretical and foundational literature to highlight specific practices that are relevant and implementable in enhancing all students' access to rich and complex thinking and reasoning skills. These include strategies for training students to notice and attend to relational correspondences, to use visual stimuli and language to draw connections, and to reduce the cognitive resources required to do so.

Keywords: relational reasoning; analogy; learning

> "*As a nation, we've reached at least a rhetorical consensus that our societal health requires us to provide all of our students with the sort of "thinking curriculum" we long reserved for a small elite—and accomplishing this will require a greater commitment to equitable education than ever before*" (Darling-Hammond, 2022)

1. Introduction

With the rapid pace of technology growth, artificial intelligence (AI), and the nature of the interconnected global economy, economic and educational aims are shifting, leading some to describe the need for Education 4.0—a new vision of education that moves away from educational norms rooted in industrial, factory-style employment to a greater focus on the ability to think critically, innovate, evaluate, and integrate ideas and fields of knowledge for new purposes (World Economic Forum, 2023). Thus, increasingly, educational goals center on cognitive skills, including the ability to think creatively, generate solutions to novel problems, weigh evidence to make judgements, and debate solutions to problems to determine most efficient and effective approaches (see OECD, 2025).

In parallel, then, employers and researchers alike have been positing that the format of schools must reflect these new goals, including updating the "factory model" of education in which discrete, topically focused tasks are given by each teacher in a series of stages throughout the day (Darling-Hammond et al., 2024). K-12 educational aims have been moving away from this emphasis to a greater focus on the abilities to think critically,

evaluate, and integrate ideas and fields of knowledge for new purposes, but what this entails for practice is not easily translated. School education standards and traditions have largely centered on clarifying the discrete series of topics that must be covered in an academic year or providing support for topically focused lesson plans.

This shift has provided new opportunities for cognitive psychological research to engage with educational practice, providing insight into how humans think, reason, and how these develop from childhood to adulthood. This article reviews research specifically on the development of relational and analogical reasoning: broadly, the ability to reason about relationships rather than the perpetual appearance of objects (Gentner, 1983). More precisely, we focus on analogy, which is a process of structured relational thinking that requires representing phenomena as systems of relationships that are then used to make inferences about the world from the relationships within and between these mental representations (Gentner et al., 2001). While seemingly abstract, the affordances of this type of reasoning apply across content areas of the traditional K-12 educational curriculum. In this review, we aim to provide novel insights into the way in which the literature on relational reasoning may inform best practices for pedagogy.

Incorporating more intentional uses of relational reasoning during instruction may be key to improving the equity and overall success of 21st century education—enabling all students to access and learn from a "thinking curriculum" (Darling-Hammond, 2022). Relational reasoning is core to innovation and creativity (Markman & Wood, 2009), as well as learning with understanding and the potential for transfer (see Gentner, 2003; Richland & Simms, 2015; Skemp, 1979).

In order to help highlight the utility and opportunities available by drawing on the relational reasoning literature to improve pedagogical practice, this article[1] will review the foundational and contemporary theories for how humans learn to reason relationally, and highlight ways that the learning context can be shaped to improve the likelihood that children can and do reason successfully. This should facilitate both learning for content and strengthening reasoning skills. We also highlight an important distinction that has emerged between *ability* to reason relationally and *likelihood* of doing so; this distinction means that, in order to lead children to leave school as thinkers who will actively engage with and reason adaptively with their environment, schools not only need to teach children to improve their reasoning skills but also must socialize them to seek and use opportunities to do so (Richland & Simms, 2015).

2. Defining Relational Reasoning for Educational Contexts

When a child is able to point to a location on a tree if asked "If a tree had a knee, where would it be?" (Gentner, 1977a, 1977b), they are doing something incredibly profound and challenging. Though this comes so naturally to humans that it sometimes goes underappreciated, such an action requires a mental representation of a human to be made, for the person to draw on stored knowledge of the term "knee", and for them to conceptualize the relative location of a knee to the rest of the body. Then, they must align that representation to something that looks different, is spatially distinct, that does not functionally use a "knee" and indeed has no mobility, and likely is not something they have ever thought of together with a human body before. After doing both, the child must take their mental representation of this spatial relationship within the human to make inferences about the relatively same spatial location between a "knee" and the full height of the tree. Thus, this process of drawing on relational similarity between these representations includes multiple steps.

Reasoning with higher-order relations involves first including the construction of relational representations (of a human/knee and a tree), finding alignment and relational

https://doi.org/10.3390/bs16020210

similarity between representations, mapping between the relational correspondences, and making inferences based on these correspondences (Holyoak & Thagard, 1989; Gentner, 1983). This can be formalized as identifying a relationship or system of relationships with representation 1, "A:B", and aligning these to the similar relationship or system of relationships within representation 2, "C:D."

A useful connection between the basic research literature on relational reasoning and more contemporary studies of classroom learning has been their highlighting the role of higher-order thinking in relational reasoning. Higher-order thinking is a potent mode of learning (see Frausel et al., 2020; Lewis & Smith, 1993), characterized by structured relational reasoning in which a learner draws a higher-order relationship between two lower order relationships or systems of relationships (Gentner, 1983). This might be "same", which means we can take what we know about one situation and apply it to another context—which would be a classic analogy. This type of higher-order thinking in a school setting would allow for instruction about one type of math problem to be transferred to solve a second math problem. Or, it could allow a student to examine two primary documents in history to recognize that the same pattern of a phenomenon is happening in both contexts (see Richland & Simms, 2015, for further discussion).

Reasoning that centers on inferences derived from finding the same higher-order relation between these relational correspondences is generally defined as analogical reasoning, with the formalization depicted as the relations between A and B are equivalent to those between c and d. A:B = C:D (Gentner, 1983). Though, empirical research centered in everyday contexts has demonstrated that, when executed in a real-world context, unless the two representations are identical, there are also differences inherent in these relational alignments—thus, in order to effectively align a spatial relationship such as "knee" across a human and tree, one must also consider and grapple with the differences between a tree and a human, such as the overall size and movement patterns (see Richland et al., 2004). This can be true when scientists try to interpret unexpected research findings (Dunbar, 1997) or when mathematics students aim to distinguish between problems that appear similar but have important differences (e.g., between a solvable inequality and one with no solution).

Cognitively, in order to conduct higher-order relational reasoning, one must also find and create mental representations of the lower-order relations (e.g., A:B; see Doumas & Hummel, 2013). These lower-order relations may involve simple relations, such as "bigger than". These simple relations often mask larger relational systems, however, as one can see if we consider a causality relation. One may infer and state that "A causes B", and then, upon further reflection, it may become clear that this is actually a higher-order relation between two relational systems: "A" here might comprise a system of relations that together cause "B." For example, the statement that "climate change may cause reduced rainfall" involves a complex system of relationships causing another complex system of relationships.

So, for the current purposes, we will focus on reviewing the literature on the ways in which children represent and reason about higher-order relations, and the implications of that work for supporting education; however, we see this as an overarching reasoning category that includes other levels and aspects of reasoning. Other forms of relational reasoning include making inferences other than structural similarity equivalence from relational correspondences, such as antithesis (Alexander & The Disciplined Reading and Learning Research Laboratory, 2012).

https://doi.org/10.3390/bs16020210

2.1. Structure Mapping in Higher-Order Thinking

The structure mapping model of analogy provides a framework for considering the cognition involved in reasoning relationally, which is highly informative for conceptualizing how to best support children in acquiring these skills, as well as to develop and test actionable pedagogical practices that align with extant learning data. Structure mapping refers to the process of creating a mental representation of relationships within a source and target object and aligning these representations in order to make inferences (Gentner, 1983). For example, if one is learning about patterns in climate change, one might make a mental representation of the relations within the range of temperatures in Phoenix, Arizona, and those in Chicago, Illinois, to find that both are experiencing a broader range than in prior decades, though the ranges themselves are different. Making this alignment allows one to make a broader inference that climate change is yielding broadening temperature swings across ecosystems.

Importantly, however, it is also the case that the pragmatics of one's larger goal-oriented reason for making structural alignments may lead to different relational representations of a source and target, and also different alignments and mappings between the representations, as highlighted by the Multi-Constraint model of Analogy (Holyoak & Thagard, 1989; Spellman & Holyoak, 1992). For example, a reasoner might represent the same range of temperature values in the same two cities described above as both demonstrating higher and lower temperatures than in previous decades, thus leading to the inference that climate change is not yielding broader global warming.

Thus, it is crucial to understand that, in order to use relational reasoning effectively in educational contexts, instructors must recognize that learners will not always create the same mental representations intended by the instructor who by definition is more of an expert in the field, and the pragmatics and pedagogical context of the opportunities for reasoning must be crafted intentionally (Simms & Richland, 2019). If not, there are ample opportunities for producing misconceptions or misunderstandings by learners (Zook & Di Vesta, 1991).

2.2. Aligning a Structure Mapping Model of Relational Reasoning with Educationally Defined Higher-Order Thinking

While the structure mapping model has generated insights for educational practice which we will review below, they are often very theoretically precise but use different vocabulary and meaning than those used by educators. Thus, we draw connections between this literature and principles of "higher-order thinking," defined as forms of thinking that necessarily involve inferences that go beyond the given information, building adequate representations, and analyzing and constructing relations (Lewis & Smith, 1993; Resnick, 1987). A common feature of higher-order cognitive functions is that they are invoked in novel situations where the simple recall or direct application of existing mental structures—typically associated with lower-order thinking—no longer suffices to achieve the desired goal (Lewis & Smith, 1993). The authors elaborated, "higher-order thinking occurs when a person takes new information and information stored in memory and interrelates and/or rearranges and extends this information to achieve a purpose or find possible answers in perplexing situations (p. 28)." To bridge the gap between the "unknown" (i.e., the desired goal or the solution to the novel problem) and the "known" (i.e., existing knowledge structure), individuals often rely on structural mapping, the core mechanism underlying relational reasoning. Through this process, learners seek both similarities and differences at the structural level in order to adapt prior solutions to novel problems.

Consistent with this view, Bloom's taxonomy of educational objectives (Anderson & Krathwohl, 2001; Bloom, 1956) emphasizes the importance of measuring learning through

https://doi.org/10.3390/bs16020210

cognitively demanding tasks (higher order thinking tasks) such as application, analysis, synthesis, and evaluation. Importantly, all of these involve cognitive processes of relational thinking Viewing learning content as coherent yet decomposable systems of relations, which is the first step or prerequisite of structural mapping, allows for insight into the cognition underpinning many of these types of tasks. For example, to apply knowledge to novel math problems and to analyze, synthesize, and evaluate math problem solutions, students must conceptualize mathematics as a system of quantitative relations governed by reasoning, rather than isolated principles to be memorized. Moreover, mapping these relations across structurally similar systems can promote creative problem solving and flexible transfer, both of which are hallmarks of deep conceptual understanding. In this way, structural mapping functions not only as a mechanism of relational reasoning but also as a foundational driver of broader higher-order thinking skills.

Importantly, these have different functions as well, with some uses of relational reasoning improving students' learning for categories or concepts, while others help to produce problem-solutions and to generate creative innovations to meet real-world problems. We will next clarify different functions for relational reasoning that are relevant to educational contexts, followed by a review of literature demonstrating that there may be differences in individuals' mindsets that lead them to notice or attend specifically to relations, which may in turn have implications for future learning quality (Zhao et al., 2025a). Finally we provide a description of specific practices that leverage the literature and that are relevant and implementable to encourage relational mindsets and structure-mapping processes in everyday learning settings. These include pedagogical strategies for training students to notice and attend to relational correspondences, including the use of relational language, and practices that instruct by using visual stimuli and language, as well as to reduce the cognitive resources required for students to notice and use relational correspondences.

2.3. Key Functions for Relational Reasoning in Education

Structure mapping is an underpinning mechanism for aligning and drawing inferences based on comparing representations, but the ultimate learning benefits of these relational reasoning instances can be quite varied. We describe two key functions of relational reasoning in educationally relevant learning: acquiring relational categories and using relational reasoning for problem solving and transfer.

2.3.1. Relational Category Acquisition

Relational reasoning underlies the human ability to understand and infer relational concepts or categories (Kalra & Richland, 2022), which are concepts defined not by any specific property of individual items but by the network of relationships that they are part of. For example, words such as "neighbor" or "barrier" derive their meaning from the roles they play within a broader structure or the interactions they have with other components. Moreover, words like "mutually exclusive," "opposite," and "homogenous" directly describe the abstract relation between two or more objects. This contrasts with feature-based categories, like "apple," "cat," or "chair," which are primarily defined by their intrinsic, observable attributes, such as shape, color, or function.

One challenge (and opportunity) in learning relational categories is that these words can be used to describe various entities that do not share salient similarity in the surface-level features. A "neighbor" can be a young lady who lives next to your house, or two nations that are adjacent to each other. Similarly, a "barrier" can be a fence that stops you from entering a house, or a challenging situation that prevents you from being promoted. To learn these relational concepts or categories, or to engage in relational reasoning, children need to move beyond the obvious, observable dissimilar surface-level features, and instead

https://doi.org/10.3390/bs16020210

engage in structural comparisons where they need to juxtapose and hold at least two sets of entities in their mind (e.g., two houses next to each other and two nations adjacent to each other), align them based on the role that each entity plays within each set, and abstract or infer the commonality in the relation that the two entities form within each set. Finding the higher-order "same" relationship between two instances of a lower-order relation, i.e., "adjacent" or "neighbor," is a potent way to generalize that the relation can apply to other contexts that share the same relational properties, (i.e., two families living next to each other or two nations adjacent to each other). The work of finding higher-order relations, or "relations among relations," can support abstraction and category learning (Gick & Holyoak, 1983; also see Alexander & The Disciplined Reading and Learning Research Laboratory, 2012; Dumas et al., 2013).

The process of noticing and learning relational concepts is more cognitively demanding than noticing a single shared property (e.g., all cats are small, furry, four-legged animals). In contrast, relational concept acquisition requires abstraction, which is effortful but at the same time provides the opportunity to apply these concepts to a broader range of contexts. Such alignment is hypothesized to derive from aligning multiple representations of relations, inhibiting attention to features that are not alignable (e.g., humans versus geographical regions) to abstract out the relations that are common (e.g., "adjacent") (see Gick & Holyoak, 1983). Once abstracted, this schema becomes easier to apply to new contexts or problem-solving situations. Also importantly there may be single levels of complexity to these correspondences and relations (two adjacent nations) or there may be increasing levels of complexity, which increase the working memory demand as well as the level of complexity of the distraction (e.g., a higher-order relation "same" may involve mapping a single level of complexity—as in "two objects are adjacent" or they can have more levels of complexity that cannot be subsumed transitively or otherwise, as in "one object is to the right and left of another object" (Halford et al., 1998).

2.3.2. Problem Solving and Transfer

Finding problem solutions by drawing on relational correspondences between representations has been recognized as foundational for higher-order cognitive functions, finding problem solutions (Halford & Andrews, 2006; Gick & Holyoak, 1980), creative innovation (Chan & Schunn, 2015; Dumas, 2018), and knowledge transfer from one context or problem to another (Gentner et al., 1993; Goldwater & Schalk, 2016; Richland & Simms, 2015) across a wide range of academic disciplines, especially including Science, Technology, Engineering, and Mathematics (STEM) (i.e., Dunbar, 1995; Farrington-Flint et al., 2007; Klein et al., 2007; Gentner et al., 2001; Zhao et al., 2021). The key role of relational reasoning in solving problems involves finding relational correspondences between representations and applying it to a new context. The reasoner must construct a mental representation of a problem or a prior exemplar in which they consider both the key features and their relationships (e.g., that one object cuts another into two). Then the reasoner aligns these two representations (e.g., both had one object cut another into two). The objects are alike, as are the "cut" relations (Rattermann & Gentner, 1998). This alignment enables several steps of reasoning: it provides the opportunity for drawing inferences about this new context by allowing one to better understand key features of both the context and the concept, to draw on what happened in one context to determine steps to provide a solution in the second, or to otherwise reorganize one's mental representation of a new context (see Goldstone et al., 1991). Moreover, core mechanisms of relational reasoning, such as structural alignment and relational abstraction support conceptual understanding and transfer in educational contexts. These mechanisms allow learners to discern relational correspondences across representations, potentially promoting conceptual differentiation

by highlighting non-alignable features and, in some cases, prompting conceptual change. Relational abstraction further allows learners to extract schemas that can generalize beyond surface features, thereby laying the foundation for transfer. Consequently, instructional designs that incorporate carefully aligned, multiple comparisons are particularly effective for fostering conceptual change and transfer. In contrast, single, highly concrete analogies may support initial sensemaking but carry a greater risk of misunderstanding and are often less effective for promoting flexible, generalizable understanding (see more discussion in the Limitations and Future Directions Section).

We next turn to elucidating the centrality of relational reasoning to education as an endeavor.

3. Learning by and for Relational Reasoning

Relational reasoning as a powerful cognitive tool can contribute to a deeper understanding of key content knowledge in many academic subjects, especially STEM. Beyond its role as a learning strategy, deliberately engaging in relational reasoning—both within and across domains—should itself be an educational objective, because it cultivates expert-like thinking and supports creative problem solving.

3.1. Relational Reasoning as a Tool for Learning Domain Knowledge

Since many concepts in STEM or social sciences are inherently relational in nature, students need to attend to and reason about the relations among the concepts to achieve deep understanding. For example, reasoning about relations is fundamental to understanding key mathematical concepts, such as fractions, decimals, and proportions (e.g., Begolli & Richland, 2016; DeWolf et al., 2016). For instance, when learning proportion, students need to focus on the higher-order relation between the two sets of objects and realize that the quantitative relation within each set of objects is the same across them (e.g., it takes 0.6 cups of flour for every strawberry to make a certain flavored cake, regardless of whether it is for a small cake which uses five strawberries and 3 cups of flour or a big cake which uses fifteen strawberries and 9 cups of flour). Relational reasoning plays a similarly important role in science learning. Many key concepts in science are grounded in abstract relations. For example, the concept of force describes and analyzes how two or more entities interact with each other. Being aware of these insights, the Next Generation Science Standards (NGSS Lead States, 2013) aimed to reshape science instruction so that it focused on interconnected systems and relations rather than on discrete facts or topics. Considerable empirical evidence also supports these propositions that relational reasoning ability is closely associated with academic performance in various domains, such as reading (Goswami & Mead, 1992), mathematics (Gentner et al., 2001; Borriello et al., 2023; White, 1998), engineering (Chan & Schunn, 2015; Dumas et al., 2016), and science (Braasch & Goldman, 2010; P. K. Murphy et al., 2017).

3.2. Relational Reasoning as an Educational Goal on Its Own

In addition to serving as a powerful learning technique, relational reasoning is a core educational goal on its own. Expertise in many domains develops through a shift from attending to surface-level features to recognizing and reasoning based on deeper relational patterns (Chi et al., 1981; Goldwater & Gentner, 2015; Rottman et al., 2012). For instance, Chi and colleagues (Chi et al., 1981) examined how experts (i.e., advanced PhD students studying physics or physicists) and novices (i.e., undergraduate students studying physics) categorize and represent physics problems through four experiments involving problem sorting and verbal protocols (Figure 1). In these experiments, participants were asked to categorize or think aloud the approach to solving a series of physics problems. These

https://doi.org/10.3390/bs16020210

problems can either be classified by surface-level similarity (e.g., problem format or theme) or structural similarity (e.g., the underlying principles shared across problems despite different problem properties). The findings consistently demonstrated that experts categorize problems based on underlying physics principles or deep relational structures (e.g., Conservation of Energy or Newton's Second Law) that guide their solution approaches; in contrast, novices categorize the problems based on superficial features mentioned in the problem (e.g., force problem, rotational motion problem). The study highlights that the ability to perceive the deep structure from the surface features is a hallmark of expertise in physics problem-solving.

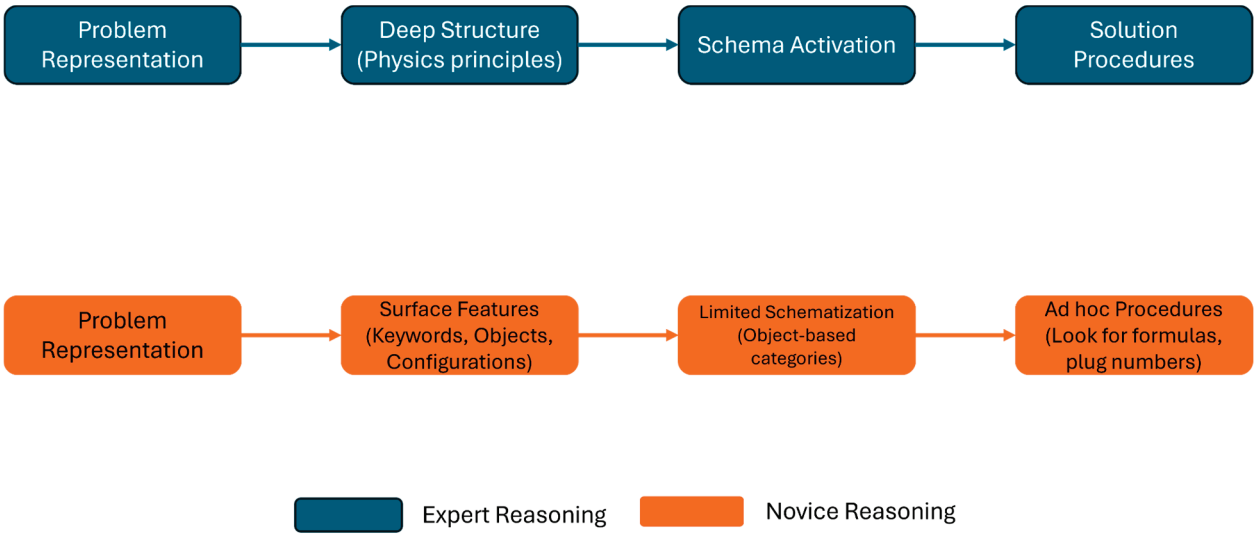


Figure 1. The steps of expert reasoning vs. novice reasoning (Chi et al., 1981).

Goldwater and Gentner (2015) further investigated the processes underlying one aspect of expertise: the ability to detect causal relational patterns across contexts that vary in surface-level characteristics (i.e., common effect, common cause, positive feedback, and causal chain in biology, economics, environmental science, and mechanical engineering). The researchers conducted experiments in which 97 college students learned about specific causal patterns through examples, varying whether students received direct instruction about the causal patterns or engaged in alignment by comparing pairs of examples across contexts. The researchers used an ambiguous sorting task consisting of 16 example phenomena, representing a matrix of the 4 causal systems crossed with the 4 content domains, to measure how participants learned causal system categories when they competed with domain information. The findings indicated that, while directed instruction improved the understanding of individual examples, analogical comparison fostered the abstraction of the common causal pattern. Crucially, the combination of both explication and alignment led to the greatest gains in the sensitivity to the causal patterns in new examples across domains. This finding also aligns with evidence that expert geoscientists spontaneously generate analogies grounded in key causal principles, reinforcing the idea that sensitivity to relational structure characterizes expert cognition (Goldwater et al., 2021). For more discussion of the centrality of relational thought to education in disciplines including science, math, and even history, see Richland and Simms (2015).

In part, the importance of relational reasoning across disciplines pertains to the fact that relational reasoning underlies the development of flexible knowledge transfer and creative problem solving. This is increasingly central to the 21st century educational mission, with national reports proposing that transforming academic content and transferring knowledge across contexts are not only the key to academic competency but also to career opportunities

https://doi.org/10.3390/bs16020210

in the 21st century (National Research Council, 2012; U.S. Department of Education, 2023). Persistent difficulties in transferring knowledge to novel situations can often be traced to failures in relational categorization: learners overlook the deep structural commonalities linking superficially different problems (Goldwater & Schalk, 2016). Elevating relational reasoning, therefore, offers a promising remedy. When instruction invites students to refine, manipulate, and interconnect their knowledge instead of treating it as fixed, static, and isolated, they develop a more flexible and well-retained cognitive network that supports effective application (Richland & Simms, 2015).

In addition, relational reasoning also plays an important role in developing and assessing creativity. Creativity is defined as a process in which individuals generate ideas and solutions that are both novel and useful (Dumas et al., 2021; Runco & Jaeger, 2012). The novelty dimension of creativity is represented as the degree to which an idea or solution diverges from conventional ideas or established solutions to a problem, which is sometimes operationalized as semantic distance (Dumas et al., 2021; Kalra & Richland, 2022). Highly creative thoughts or solutions may share fewer perceptual features with conventional ones, reflecting their novelty, but may potentially share the same or similar relational structures, ensuring that they are considered useful. Recent research studies have also investigated and supported the link between creativity and relational reasoning (Chan & Schunn, 2015; Dumas, 2018; Dumas et al., 2021; Green et al., 2012; Vartanian, 2012).

Chan and Schunn (2015) observed a team of 10 professionals from a range of design-related disciplines, such as mechanical engineering, electronics and business development, and industrial design; they were observed in a naturalistic setting when they brainstormed conversations for developing a new product concept for a new market. The researchers examined whether and how relational reasoning, specifically far analogies, led to the generation of creative ideas. Researchers recorded and transcribed two group meetings. Analogy (i.e., near or far) was identified and coded when a designer referred to another source of knowledge and attempted to transfer concepts from that source to the target domain. Near-analogy involved mappings from sources that were related to tools, mechanisms, and processes associated with graphical production and printing, whereas far-analogy referred to mappings from sources which were more distant from the subject at hand. The outcome variable of interest was the novelty of concept generation, which was operationalized as functional distance: the degree to which a given concept's described functionality was different from a prior concept, rated on a scale ranging from 1 to 5 with 1 being very similar and 5 radically different. They found that far analogies contributed to novel ideas through detailed and focused exploration of the discussed ideas related to the problem, which leads to new and useful versions of the existing concepts in an iterative way.

Dumas (2018) assessed relational reasoning ability and divergent thinking among 77 college students. He used the Test of Relational Reasoning (Alexander et al., 2015), a visual–spatial measure with 32 items in total tapping on four forms of relational reasoning, and the Alternate Uses Task to assess divergent thinking, in which participants were asked to produce as many novel uses for a given object as possible (Guilford, 1967). Two dimensions of divergent thinking were scored—fluency and originality. Fluency was operationalized as the number of uses generated by each participant for each object, and originality was scored as the semantic distance between each participant-generated use (e.g., read; throw like a Frisbee) and the object-prompt (e.g., book) using Latent Semantic Analysis. Quantile regression analysis was then conducted to explore the potentially different predictive effects of relational reasoning ability at varying levels of divergent thinking. The study found that relational reasoning ability significantly predicted originality, a key aspect of divergent thinking, particularly for students scoring at or below the median of originality. Relational reasoning was found to significantly predict fluency

https://doi.org/10.3390/bs16020210

only for children scoring at the 90th percentile, implying that the association between relational reasoning and fluency might be confined to those who produce many ideas on the Alternative Uses Task. Therefore, relational reasoning may be an important process or tool in the development of creativity.

Taken together, teaching students how to engage in relational reasoning and abstract relational structures underlying various contexts that differ in surface-level features should be a key educational goal.

4. Best Practices for Teaching by Relational Reasoning

Beyond training youth to conceptualize their learning as relational inquiry, designing relational reasoning opportunities for teaching content can have strong implications for teaching practice so that encoded and stored constructs are conceptual, meaningfully embedded in prior knowledge structures, and transferable to new contexts. We review the literature on this topic, highlighting the key ways in which teaching can be supported through analogy or relational reasoning opportunities.

4.1. Using Analogy to Fuel Understanding of Abstractions

Analogy is the mental alignment of the systems of relations within two representations, such as between the relationships that govern electric circuits and city water system. For a classic analogy, the higher-order relationship between the two systems of relations (here both "pumped by pressure and circulate") is "same" (Gentner, 1983). The act of making such analogies allows the reasoner to not only be able to explain the specific representations being elaborated, but also it can fuel an abstraction process in which the reasoner recognizes the common pattern to both representations. This learning mechanism was identified in early problem-solving studies (Gick & Holyoak, 1983), but further research has highlighted how it can be leveraged to teach real-world classroom materials (See Richland & Simms, 2015; Vendetti et al., 2015). In history, for example, a teacher might instruct her class to notice how the U.S. Tea Party rebellion was akin to Native Americans' attempts to resist U.S. colonization in the Battle of Wounded Knee. Importantly, such analogies not only lead to identification of similarities—where both are actions of a colonized people seeking to separate from the economic and literal domination exerted by another society—but may also highlight key differences. The short- and long-term outcomes in the two battles were strikingly different, with the Tea Party being non-violent, and the Battle of Wounded Knee being a massacre. Further, U.S. colonists ultimately threw off the mantle of the British colonizers, while the Native Americans' sovereignty was compromised, entering a unique status of protected sovereignty subsumed by the United States government. After reviewing the two examples of colonial resistance, if a student also learned about resistance acts conducted in India and Brazil, she might induce and store an abstraction about colonialism. Perhaps, understanding what colonialism means—one nation conquering another—or drawing the developmental insight that communities who have been led by a conquering nation might be forced to fight back to reclaim their sovereignty. This process of abstraction leads to greater flexibility of that insight or concept (see Gick & Holyoak, 1983), meaning it is more likely to be retrieved when the surface features of the original source and target items are not present in a new target context. If this same student was next to be introduced to the context of South Africa, for example, she might recognize that this is again a colonial context, despite being on a different continent, involving distinct countries and cultures, and engaging different processes of resistance.

Thus, instruction by analogy can lead to several different opportunities for building knowledge. First, these rich comparisons yield stored knowledge representations that are embedded in a network of meaningful higher-order relationships (see Richland & Simms,

https://doi.org/10.3390/bs16020210

2015 for further discussion). By doing so, the likelihood that a learner will be able to recall instructed representations is higher, and the likelihood of recognizing relations to related information is also increased (Gick & Holyoak, 1980, 1983). This can be affected by the effort of the analogy-making. One dimension is the distance of the relationship between the source and target analog. Near analogies typically share some surface features or domain context between source and target: for example, two rate problems in mathematics, with one that uses the context of driving and one that uses the context of bicycling. Far analogies typically do not share surface features and may cross domains, such as between an electric current and water in a garden hose.

4.1.1. Practice: Supported Use of Far-Distance Analogies

Halpern and colleagues (Halpern et al., 1990) demonstrated that the distance of an analogy can impact retention with greater effort leading to greater retention. They found far analogy aided understanding and retrieval of scientific information. In that study, 193 adult participants between the ages of 17 and 64 read three texts on the topics of the lymphatic system, enzymes, and electricity, each with either a near analogy, (e.g., comparing the pull of gravity to electricity flowing through an electrical circuit), a far analogy, (e.g., comparing electric current to water in a garden hose), or no analogy. All combinations of near, far, and no analogy appeared equally often in combination with the three topics. Immediately after reading the text and one week later, participants were assessed on the information retained by four types of tests: (a) a free recall where participants wrote as much as they could remember about each text; (b) a cued recall, in which participants answered the prompted questions (e.g., "Why are metals good conductors of electricity?"); (c) a new analogy item based on the information provided in the text (e.g., "Hydrolysis is to synthesis as scissors is to (1) knife (2) organic solvent (3) acid (4) stapler"); (d) two inference questions for information that could be inferred but not explicitly stated (e.g., "What might happen to the lymph flow if a person were paralyzed?"). The results of a series of ANOVA tests indicated that individuals who read the scientific text with a far analogy outperformed the other two conditions (i.e., scientific text with near analogy or no analogy) on almost all the four tests; that is, they recalled more information in the free recall and cued recall tests, and correctly answered more questions that required inferences or comprehension of a novel analogy both at the immediate and delayed posttests (nonsignificant difference was only found for the solution of new analogy). From an educational perspective, teachers may consider including far analogy examples or practice for students to make efforts in establishing connections across seemingly distant domains for better knowledge retention.

4.1.2. Practice: Teaching by Highlighting Alignable Differences

Such alignable differences have been demonstrated to be particularly salient when all other aspects of the aligned representations share salient features. Kok and colleagues (Kok et al., 2013) demonstrated this in the context of reading X-rays, a notoriously difficult skill due to the many aspects of these visual representations to attend to simultaneously. Though identifying abnormalities within X-rays is challenging, when presented with two X-rays simultaneously that were aligned in all surface features aside from a key difference of physiological importance, participants easily identified discrepancies (see Matlen et al., 2020).

Similarly, when teachers aim to highlight a key difference across representations (e.g., to explain how an underlying mechanism works), learners benefit from providing visual representations that use perceptual similarities to highlight the similarities and alignments. Doing so makes discrepancies especially visible, allowing for the key alignable differences between the representations to be clarified. Hansen and Richland (2020) tested this approach in the instruction of mitosis and meiosis. They provided students with visual

https://doi.org/10.3390/bs16020210

representations of mitosis and meiosis, which appeared to be identical and showed the same steps for cellular replication except for illustrating that meiosis includes an additional step in which the cells are split into gametes with half of the total chromosomes. When students saw these two representations lined up on the same page, they were quick to point out the difference between the constructs. This was more effective than having them view and engage with the representations one at a time in sequence, despite the fact that this would likely have reduced their cognitive load to interpret complex diagrams (Mayer, 2014).

Thus, teaching by analogy by providing carefully curated source and target representations to compare can be a potent tool for encouraging abstraction and process learning, in addition to storing and retaining the instructed representations. To use the colonialism analogy begun above, this may lead to improved encoding of both the specific instances that were instructed—here the Boston Tea Party and the Battle of Wounded Knee—and the abstraction that renegotiating colonialist relationships may require uprising. Doing so using representations with similar features but that possess easily aligned relational differences may allow those aspects of the systems that differ in meaningful ways to be noticed.

4.1.3. Explicit Instruction for Students to Use Relational Correspondences

While analogy is potent, it is not the only form of relational comparisons that can be useful with higher-order relations other than "same" being possible across representations. Two relationally complex structures may be causally related, or hierarchically structured, for example. Additionally, the structures may have similar conceptual structure though the details of their representations may not be easily aligned. Crucial to ensuring this is effective for learning, however, is making sure that students do the intended relational reasoning.

A productive and informative study that examined learning from comparisons between mathematical solution strategies demonstrates this instructional tool and opportunities. Mathematics teachers regularly use comparisons to instruct in mathematics (Richland et al., 2006, 2007), perhaps due to the crucial nature of mathematics instruction to teach abstraction, transfer, and flexibility—features that, as described above, are particularly well taught by analogy.

It is more effective to create learning opportunities in which students are required to make comparisons and alignments. In one set of studies, pairs of students were asked to verbally compare and contrast standard and non-standard solution strategies to algebra problems (Rittle-Johnson et al., 2009, 2017; Rittle-Johnson & Star, 2007; Star & Rittle-Johnson, 2008). These opportunities led to gains in procedural knowledge, flexibility, and conceptual knowledge, provided students had sufficient prior knowledge of the key content (Rittle-Johnson et al., 2009; Rittle-Johnson & Star, 2007; Star & Rittle-Johnson, 2008). For example, Rittle-Johnson and Star (2007) randomly assigned seventy seventh-graders to two conditions. In the comparison condition (n = 18 pairs), students studied sets of two strategies for solving the same algebraic equation side-by-side in worked problems and discussed their differences. The other condition was the sequential condition (n = 17 pairs), where students studied the same two strategies but in two different worked problems presented sequentially, and then they were asked to reflect on one single problem.

Students worked in pairs over two days, solving practice problems and writing explanations. Results of a series of multilevel modeling analyses showed that students in the comparison group made significantly greater gains in procedural knowledge and flexibility, including increased use of more efficient, non-conventional methods. Therefore, comparisons that highlight the structural differences can be a powerful instructional tool to enhance learning.

https://doi.org/10.3390/bs16020210

4.1.4. Generating Explanations and Self-Explanations

Generating explanations and self-explanations are another paradigm that leads to relational construction (Atkinson et al., 2000; Chi et al., 1994; Lombrozo, 2006). Explanations are inherently causal, meaning that children who are asked to explain their thinking, either explaining a process or phenomenon to others, or to themselves via self-explanation, are typically engaging in causal construction of their knowledge, which can develop relational stored representation and lead to both conceptual knowledge and transfer. During the explanation process, children have an opportunity to integrate new material with prior belief, identify gaps in knowledge and update it to accommodate new information, and seek broader and unifying generalizations available for transfer in order to understand the new information that they are trying to explain (see Lombrozo, 2012 for a review). Chi and colleagues (Chi et al., 1994) developed the paradigm of self-explanation, assessing the effect of self-explanation on learning among 24 eighth-grade students. As an assessment of prior knowledge, all students were asked a set of questions about each of the 23 terms on human circulatory system and to draw the path of blood flow throughout the body before the experiment started. Fourteen participants were then asked to read a 101-sentence passage that introduced how the circulatory system works in human beings. After reading each sentence which was printed on a separate piece of paper, these participants were prompted to explain what each sentence means. In addition to these general prompts, they were also prompted to explain the function of 22 key components of the circulatory system, such as atrium. Another ten participants in the control condition were not prompted for either self-explanations or functions. Instead, they were asked to read the same passage twice, which took approximately the same time on average as the total time that participants in the experimental condition took. All students were then asked to answer the same set of questions about the 23 terms on human circulatory system and redraw the path of blood flow. Results showed that the prompted group had a greater gain from the pretest to the posttest compared to the control group. Moreover, students who generated a large number of self-explanations in the prompted group (n = 4; high explainers) had a better understanding than low explainers (n = 4). Additionally, researchers inferred and drew the mental model of circulatory system for each participant based on their answers and blood path drawing. They found that all the high explainers achieved the correct mental model, whereas many of the unprompted students and low explainers did not.

A large body of research has built on the self-explanation effect paradigm, with much of the work demonstrating that explanations can lead to relational alignment and structure mapping, ultimately leading to learning and abstraction (Lombrozo, 2012).

4.2. Challenges and Practices in Relational Instruction

While relational reasoning is a potent and powerful instructional practice, it is also effortful, and there are challenges that must be considered to ensure its efficacy. We describe key challenges that must be considered, with specific practices identified for meeting these needs on the part of students.

4.2.1. Noticing the Utility of Reasoning Relationally

Reasoning relationally often involves disattending to perceptual information and recognizing correspondences between representations that do not share perceptual features. Expertise in a domain leads reasoners to be more likely to notice and focus on relational correspondences (Chi et al., 1981); however, by nature, students are non-experts in the domains in which they are being instructed, so teachers may need to be very explicit about when students should be seeking relational correspondences and when they should be making relational inferences. Researchers have identified several different types of

https://doi.org/10.3390/bs16020210

educational practices that can be effective at supporting students in doing so. We describe some of these practices in turn, which include explicit cuing and structuring a learning context to set up progressive alignment.

Explicit Cuing

Providing explicit cues for learners to compare prior instructed problems with the target problems improved the likelihood that learners noticed the relations on the structural level between the two types of problems (Blessing & Ross, 1996; Gick & Holyoak, 1980; Novick, 1988; Novick & Holyoak, 1991). Those cues could be general reminding statements for comparison (e.g., "To figure out how to solve Problem 2, think about how we solved Problem 1") or specific information about the elements that can be mapped productively across problems. For example, when teaching two math word problems, one about baking a cake and one about making lemonade, a teacher might ask: "How are the eggs in the cake like the lemons in the lemonade?". The role of cues was studied in early experiments to determine whether learners' difficulty was in encoding and representing relations for transfer or recognizing the need to align representations. Gick and Holyoak (1980, 1983) demonstrated that when explicitly cued, most participants who read two problems were able to map the correspondences and learn from the relationships, so the challenge was in noticing the key relations.

Progressive Alignment

Moreover, guiding students through a series of comparisons using a strategy in which one aligns progressive alignment strategy can also draw students' attention to the structural similarity between two seemingly different types of problems (Kotovsky & Gentner, 1996). The key to the progressive alignment strategy is to start from problem sets that share more surface-level similarities as well as structural similarity, and then gradually move to problem sets that share less and less similarities on the surface level, and finally compare problem sets that only share structural similarity. For example, when teaching the multiplicative relation, $y = 1.5x$, following the progressive alignment strategy, teacher could first use two recipe problems for students to compare, in which the quantity of one ingredient (i.e., flour/water) is always 1.5 times of the other one (i.e., milk/orange). Then, teacher could use one recipe problem and one painting problem with different numbers from the first pair, in which the quantity of one thing (i.e., flour/green paint) is always 1.5 times of the other (i.e., milk/red paint) in each problem. Finally, the teacher could introduce the equation $y = 1.5x$ and encourage students to map x and y to their correspondences based on the role they play in each of these stories, thus allowing students to understand the multiplicative thinking underlying these problems.

4.2.2. Eliciting a Relational Mindset

While explicit directions to attend to relational correspondences can be an effective practice, it is also important and useful for reasoners to develop an implicit focus on relations. This can derive from an ideological understanding of what constitutes knowledge in a domain, such as creating an understanding of mathematics, science, or history that recognizes the importance of finding the causal and relational structure embedded in the content being taught.

It can also derive from a generalized attention to relationships, in which a growing body of the literature suggests that it may be an individual difference that is relatively stable, or at least a trainable attentional pattern among both adults and children. Currently, researchers have been using the spontaneous scene analogy task to measure this attentional tendency to relations (see A. N. Murphy et al., 2021, for test details). During the task, participants were presented with two pictures depicting relation(s) that were the same

https://doi.org/10.3390/bs16020210

but using different objects. Participants were instructed to "choose one of the things in the bottom picture that goes with the xxx in the top picture". They could match objects across the two pictures either based on relational similarity (i.e., the object that plays the same role as the target from a relational perspective despite looking differently) or based on object similarity (i.e., the object that looks exactly the same despite playing a different role from a relational perspective). Relational attention was measured as the degree to which participants overcame the tendency to attend to object similarity (i.e., selecting the object match), and instead spontaneously attended to the relational similarity by selecting a relational match. The picture materials controlled for number of items per scene, and analyses controlled for executive function and age—related variables that could also explain changes in children's relational thinking (Zhao et al., 2025a). Vendetti and colleagues (Vendetti et al., 2014) and Walker and colleagues (Walker et al., 2018) demonstrated that adults could be primed to notice relational correspondences by doing a challenging, high-demand relational task—generating solutions to far-distance analogies—before mapping correspondences on this spontaneous scene analogy task. In a developmental study building on this paradigm, Simms and Richland (2019) found that, like adults, young 3–4 year old children who were prompted to generate and articulate the relations between objects in a relational task by themselves were more attentive to relational similarity and less to object similarity in a subsequent task than those who passively listened to an experimenter describing the same relationships.

Moreover, increasing research also suggests that eliciting this spontaneous relational attention or relational mindset may have implications for learning and information processing. Zhao and colleagues (Zhao et al., 2025a) found empirical evidence showing that a relational mindset may have direct benefits in classroom math learning. They used a latent class analysis to explore individual differences in relational attention among 272 elementary school students. Four distinct clusters were identified: (a) Relational Attenders, who consistently attended to relational similarities; (b) Emerging Relational Attenders, who attended to relational similarities but not as consistently as the Relational Attenders; (c) Object Matchers, who consistently attended to object similarities; and (d) Inconsistent Responders, who had an inconsistent pattern in attending to similarities. The researchers then fit a mixture model to explore the potential implications of these individual differences in relational attention on students' math learning outcomes. They found that children who consistently and spontaneously attended to role-based relations in a non-mathematical relation task (i.e., Relational Attenders and Emerging Relational Attenders) had better learning outcomes from the same 20 min math lesson than their peers who consistently and spontaneously focused on object features (i.e., Object Matchers), controlling for executive functions and prior math knowledge. Therefore, having an implicit attentional pattern that prioritizes relations, may be a powerful tool for achieving optimal learning.

Relatedly, Goldwater and Jamrozik (2019) found that relational mindset could boost retrieving analogical relations from multiple texts read by undergraduate students as long as such relational mindset was elicited before the information was encoded. Following the paradigm of Vendetti et al. (2014), undergraduate participants completed either an analogy task that aimed to elicit their relational mindset or a semantic word-association task serving as a control. They then read multiple texts from different domains, each about a particular relational concept (e.g., "trade-off", "pre-emption") explored in a particular domain (e.g., medicine, computer science). Following that, they read another similar set of texts, each could be matched with an article in the previous set based on relation (e.g., both about "trade-off" but in different domains) and domain (e.g., both in medicine but about different relational concepts). The researchers then assessed the degree to which undergraduates retrieved previously read information based on relation (e.g., they recalled

https://doi.org/10.3390/bs16020210

a previously read text on "trade-off" in medicine while reading the text on "trade-off" in computer science) or domain (e.g., they recalled a previously read text in computer science but about "pre-emption" instead). They found undergraduate students who completed the analogy task prior to reading the first set of texts retrieved more information based on relations than their peers in the control condition whose attention was not primed to relation before reading, suggesting that high-quality relational encoding was very important for subsequent relational retrieval and that inducing a relational mindset could improve those encodings.

Therefore, it is reasonable to speculate that eliciting a relational mindset before learning activities take place might have implications for the quality of the key information encoded during the learning phase, thus having consequences for learning outcomes. Training young children with a warm-up activity that requires them to actively think relationally is likely to elicit a relational mindset, which might contribute to learning from the subsequent instructions. Zhao and colleagues (Zhao et al., 2025b) demonstrated that a warm-up activity that required children to actively think relationally in a non-explicitly math context successfully directed their attention to mathematical relations in a subsequent, different task. The elicitation of relational mindset also had potential in improving children's learning from a video math lesson.

4.2.3. Exposure to and Use of Relational Language

Exposure to and using the language of relations can support children in building relational reasoning skills, both through encouraging a relational mindset, as well as to support growth in relational skills over time. This has been examined both in naturally occurring everyday speech, as well as in experimental designs manipulating specific language production. We briefly review these two bodies of literature in turn.

In everyday speech, linguistic and discursive patterns can invite and invoke relational thinking in both the speaker and the listener. Close analyses of everyday home language experiences for U.S. children have revealed discursive patterns of comparison (Silvey et al., 2023) and, more broadly, relational Higher-Order Thinking Talk (HOTT, Frausel et al., 2022). HOTT involves language that is involved in representing, inquiring about, and conceptualizing the world relationally through linguistic practices. We define HOTT specifically to include discourse that links ideas to make inferences, draw comparisons and analogies, construct hierarchies and categories, invoke schemas and link definitions to abstractions. Consider a teacher who asks the following question: "Why do you think this problem ended with $x = x$, while that problem ended in $x = 2$? What does that mean about what numbers can be used to satisfy the equations?" This mathematical prompt by a Chinese teacher (Williams et al., 2009) not only draws students' attention to a distinction that provides meaningful insight into the nature of variables and the equality relationship, but also it socializes students to know that mathematics is a field that requires reasoning to question why patterns emerge.

Patterns of acquisition and usage, as well as the longer-term implications of early HOTT, were investigated in a sample of 60 families selected to reflect the demographic range of Chicago (31 girls; 34 first born children). Spontaneous home talk was videotaped while a primary caregiver was present for 90 min at a time, three times a year, from 14 months through 58 months. All of these interactions were then transcribed systematically, with high reliability calculated between transcribers (see Rowe & Goldin-Meadow, 2009). Each utterance was coded and given a score for whether it displayed or encouraged HOTT, yielding over 1 million utterances coded for analysis.

Crucially, these early HOT utterances, both child and parent, were impactful in predicting future STEM and reasoning skills. Parent HOTT was correlated with SES and

https://doi.org/10.3390/bs16020210

parent education level, which were correlated with children's HOTT onset of usage and frequency (Frausel et al., 2020). Further, children's early frequency of using HOTT that was structurally meaningful predicted children's 5th grade performance on a verbal analogies task, in addition to reasoning and inferencing on a reading comprehension task. Thus, these patterns of talk in early childhood may shape the skills that children bring to learning and to educational contexts going forward.

Analyses of these HOTT data also revealed that HOTT was invoked regularly while parents or children told personal narratives—stories about one's experiences in the past or expected experiences in the future (e.g., a planned trip to the zoo; Frausel et al., 2021). This finding suggests personal narratives may be a leverage point for increasing the amount of HOTT children are exposed to.

Relational language has also been investigated as a tool for improving reasoning on specific tasks (see Loewenstein & Gentner, 2005; Pruden et al., 2011). Loewenstein and Gentner (2005) conducted a series of experiments and found that hearing the language of spatial relations (i.e., "on", "in", "under", "top", "middle", "bottom") enhanced the spatial relational encoding, thus improving young children's spatial mapping performance. Young children who were provided with the spatial relational language during the mapping or orientation phase from the experimenter, for example, "I'm putting the winner *on/in/under* or *at the top/middle/bottom of* the box", outperformed children in the baseline condition on a series of spatial correspondence tasks, where such spatial relational language was replaced by general language (e.g., "I am putting the winner *here/there*") and accompanying gestures, after controlling for age and the location of the target object. The results were consistent across five experiments, and researchers found that the effects of such spatial language were maintained over time. Children who were provided with the spatial terms during the initial orientation phase outperformed their peers in the baseline condition when both groups carried out the mapping task two days later, with no further exposure to the spatial terms.

In another longitudinal study the same interactions that had been coded for HOTT (Frausel et al., 2021) were now coded for a specific type of relational language—spatial language (Pruden et al., 2011). In the dataset used for this analysis, researchers examined the frequency of spatial language inputs in data collected with 52 children between the ages of 14 months and 46 months in a total of nine visits, in which unstructured, spontaneous home interaction was recorded once every four months for 90 min. In addition to these naturalistic observations, researchers also evaluated children's performance on three non-verbal spatial tasks and their receptive vocabulary when they were at 54 months of age. When analyzing the video transcribed data, researchers focused on the use of language about the spatial properties and features (e.g., "circle", "triangle", "big", "little", "bent", "curvy") from parents and children separately. Considerable individual differences in the use of spatial terms were found for children and parents (i.e., 5 tokens of spatial language to 525 spatial tokens for parents, 4 to 191 spatial tokens for children over the 9 visits). Importantly, the study found that, when parents tend to use more spatial language, their child was likely to use more spatial language, and vice versa ($r = 0.55$, $p < 0.001$), after controlling for parent's overall language use. Moreover, children who produced more spatial language were also likely to perform better on spatial problem-solving tasks at a later age ($0.33 \leq rs \leq 0.39$, $p < 0.05$) after controlling for the parent's and the child's overall language use.

4.2.4. Activating Relevant Prior Knowledge

Adequate prior knowledge is necessary to facilitate and enable productive relational reasoning. The developmental literature on analogy originally derived from Piaget's

https://doi.org/10.3390/bs16020210

assertion that children were not able to reliably solve analogy problems until they had reached formal operations, typically adolescence, due to their low success on his analogy problems until that point (see Goswami, 2002 for discussion). New paradigms that took children's knowledge base into account and ensured children had strong understanding of all included relations demonstrated, in contrast, that children could reason reliably about relations by 3–4 years of age (see Goswami & Mead, 1992; Rattermann & Gentner, 1998). Therefore, adequate prior knowledge is necessary for children to engage in relational thinking, though it can be a mechanism of learning as well.

Importantly, Gentner (1988) posited the relational shift, which is the observation that, as people build knowledge in a domain, they move from noticing and reasoning on the basis of perceptual features to noticing and reasoning on the basis of relational structure. For example, this was observed in adults in a business school learning context. When asked to compare cases, students first matched by finding alignments between the surface content of the cases, while with greater knowledge, they mapped on the basis of relations (Loewenstein & Gentner, 2005).

Children also need sufficient prior knowledge to benefit from analogical comparisons. In an algebra task in which students were assigned partners and randomly assigned to conditions in which they were asked to either compare solution strategies simultaneously or to discuss them one at a time in a sequence, students' level of algebra knowledge at the introduction to the lesson was predictive. If on a pre-test the students made any attempt to solve algebra problems by moving calculations across the equal sign, showing some level of prior knowledge in solving equations, even if incorrectly, they were likely to learn better from the comparison condition than the sequential condition. When students did not begin the intervention with these early skills; however, they benefited more from learning about the problem solutions separately (Rittle-Johnson et al., 2009). Therefore, ensuring and potentially activating the relevant prior knowledge allows learners to engage in and benefit from relational instruction.

4.2.5. Alleviating Cognitive Load to Support Relational Reasoning

Relational reasoning relies heavily on executive functions, required for mentally representing relational representations, aligning and mapping, inhibiting irrelevant correspondences, and inferring structure and information about problem solutions (Krawczyk, 2012; Krawczyk et al., 2008; Morrison et al., 2004; Waltz et al., 1999). Considerable evidence from neuroscience supports the link between prefrontal cortex—an area that primarily governs including selective attention, inhibitory control, and working memory—and performance on relational reasoning tasks. Krawczyk and colleagues (Krawczyk et al., 2008) and Morrison and colleagues (Morrison et al., 2004) investigated adults with frontotemporal lobar degeneration (FTLD) and found that those with damage to the prefrontal cortex performed more poorly on relational reasoning tasks than those with damage to the temporal lobe, particularly when the tasks demanded greater inhibition of distracting information. Similarly, children with traumatic brain injury may display lower levels of relational reasoning, in part explained by damage to the executive function resources (Krawczyk et al., 2010). In young children, EF resources that are still limited by development means that children are more susceptible to inhibition failures (see Richland et al., 2006). Researchers found that children who were successful at mapping relations in scene analogies performed at chance when required to inhibit responding to a perceptually salient lure.

Educationally, these mechanisms have implications for how one might structure a pedagogical context to reduce the burden on learners, inhibiting attention to irrelevant correspondences, and to align and map key relational correspondences. The specific details of how one presents information during instruction may be crucial to predicting whether

https://doi.org/10.3390/bs16020210

learners will notice key relations, represent the information relationally in mind, and align and map relevant and crucial relational structures. Doing so in a way that reduces the cognitive load for conducting structural mapping can improve the likelihood that this is the case (Catrambone & Holyoak, 1989; Gick & Holyoak, 1983; Novick & Holyoak, 1991; Richland & McDonough, 2010; Rittle-Johnson & Star, 2007; Ross & Kennedy, 1990; Star & Rittle-Johnson, 2008). Some of the these commonly used instructional approaches include presenting the source analogs (e.g., two problems or strategies) simultaneously and making them visible together while they are being instructed (see Begolli & Richland, 2016; Kok et al., 2013), using visual alignment and linking gestures to highlight to correspondences between items being compared, and encouraging mental imagery or visualizations for comparisons (Alfieri et al., 2013; Matlen et al., 2020; Richland & McDonough, 2010).

Reducing Cognitive Load by Making Comparisons Visible: Simultaneous Versus Sequential Presentation of Representations

Viewing two representations simultaneously can result in high cognitive load due to requiring a large amount of information to organize attention to find and map relational correspondences (Mayer, 2014). Richland and McDonough (2010) conducted two experiments to assess the degree to which these instructional approaches that should facilitate structural alignment by reducing cognitive load would enhance learning and transfer. College participants learned mathematical concepts (e.g., permutation, combination) through carefully controlled video instructions that varied in the level of cuing to relational structure, including comparative gesture, source visually supported, source visible during instruction of target, visual alignment, and the opportunity for students to actively engage in judging the accuracy and mapping correspondences between different strategies. They found that the high-cuing condition resulted in higher rates of positive extension of learning to new contexts, and lower rates of susceptibility to misleading contextual features when assessed both immediately and one week later.

Further work since that laboratory-based study has examined the benefits of these practices in more ecologically valid educational contexts, including in classroom mathematics lessons on proportional reasoning (see Begolli et al., 2018 for review), as well as in science. As described above (Hansen & Richland, 2020), providing children with opportunities to learn from meiosis and mitosis representations on the same page was more effective for learning than learning from them in sequence.

In another quite different mode of instruction, in a game-based math learning paradigm, similar results were identified for teaching rational number conversion using opportunities for relational reasoning (Bustamante et al., 2022). When compared to business as usual, games that required students to compare representations and move between decimal and rational number representations led to greater learning and transfer to new rational numbers. For those who began with adequate skill, a test of simultaneous versus sequential instruction also showed similar results to those described above. Games either incorporated multiple different denominators that had to be converted to decimals on the same game play board, or they were kept separate onto different game boards. When taught together, the overall learning was higher, and there was better transfer to new number representations for the simultaneous representations, but this did depend on students' prior knowledge.

These results provide indication that these cognitive principles of analogy and relational reasoning can inform instruction productively through traditional instruction as well as more innovative approaches such as game play.

To this end, recent developments in large language models (LLMs) and AI have also led to exciting opportunities and possibilities. While generative AI does not seem to process higher-order relations using similar representational approaches to a human,

https://doi.org/10.3390/bs16020210

there has been rapid evidence for the emergence of relational reasoning capabilities in these systems. This indicates this may be a productive way to support relational reasoning in academic contexts in youth and learners as well, since engaging learners in HOTT or relational reasoning requires the skill to participate in this level of thought as well. Webb and colleagues (Webb et al., 2023) demonstrated that models such as GPT-3 and GPT-4 can perform at, and in some cases even above, human levels on a variety of analogical tasks, including problems structurally similar to Raven's Standard Progressive Matrices. Lu et al. (2019) also developed computational models capable of generating analogical mappings from initially nonrelational inputs. Theoretically these advances suggest there are different pathways for conducting analogical reasoning, since LLMs may not perform relational representation as a human might, but most importantly for the current manuscript, this work indicates the potential for computational systems to generate and develop analogical reasoning opportunities. These developments suggest that AI has the potential to become a valuable aid for relational instruction. For example, teachers might use AI to generate problem sets that differ in surface features but share deep structural similarities, allowing students to compare them. Similarly, students could engage AI tools to explain analogical mappings step by step, deepening their understanding of the underlying relations.

5. Limitations and Future Directions

Overall, the deep cognitive literature on relational reasoning has great potential for guiding and contributing to instructional practice in several key educational domains, yielding precisely the type of flexible, generalizable learning that is increasingly important to the demands of the 21st century.

Although we have emphasized the utility and benefits of engaging in relational reasoning for learning, it is also important to acknowledge its limitation. In particular, analogical reasoning can sometimes mislead learners away from the intended learning goals to forming misconceptions, especially when learning complex concepts. As highlighted by Spiro and colleagues (Spiro et al., 1988), exact one-to-one mappings between source and target domains are rare (i.e., all the key relations in the source can be mapped one-on-one to all key relations in the target); more often, only a subset of key elements in the target aligns with the source. While such partial mappings may help novices gain an initial understanding of the target, they risk producing incomplete or even erroneous conceptions. Misunderstanding can arise when the source domain of an analogy is inadequate or potentially misleading for understanding the target domain. In such cases, learner's understanding may be reduced to only those features explicitly mapped from the source, leading to oversimplified or distorted representations of the target concept. For example, when household plumbing is used as an analogy for teaching blood flow in human body, learners may focus on resistance as the sole cause of impedance in blood flow while overlooking critical nonaligned properties, such as the flexibility of blood vessels in contrast to the rigidity of pipes. As a result, students may struggle to understand or may entirely neglect additional sources of impedance in blook flow, such as compliant and inertial reactance, which play essential roles alongside resistance. This tendency to reduce complex systems to the core relations highlighted by a single analogy exemplifies how analogical instruction, if not carefully scaffolded, can foster entrenched misconceptions. Moreover, these oversimplified representations can be resistant to change, persisting even as learners progress to more advanced stages where a deeper understanding of the target's complexity is required. In addition, teaching by analogy runs the risk that young learners may focus on surface features or irrelevant relations when mapping across domains. Zook and Di Vesta (1991) similarly found that children who learned through analogy generated more

https://doi.org/10.3390/bs16020210

incorrect inferences from surface-level similarities or irrelevant features than those who were not made aware of the analogy, consistent with the relational shift hypothesis.

Thus, when using analogy in instruction, particularly with young learners, teachers should anticipate where misalignments may occur and proactively mitigate them through explicit guidance.

6. Discussion and Implications

This study therefore indicates a set of applications for educational practice that are both important directions for further research, bridging the cognitive science of relational reasoning and education, and that are usable principles for informing practice. We summarize these here as an opportunity to consider direct ways to use this research for instruction and encourage researchers to test their relevance across disciplinary foci. Though teachers who are not attuned to the use of comparison may not adopt these practices easily (see Star et al., 2015), training that incorporates teacher professional development along with curriculum materials designed to clarify meaningful comparison and contrast opportunities may be a productive way to support adoption of these practices (see Durkin et al., 2023).

1. ***Ensure a learner recognizes the utility of relational reasoning:*** This may take the form of directed attention with speech that highlights the usefulness of referencing a prior referent, use of relational language, or drawing attention to one-on-one relational correspondences within representations. This can also take the form of using spatial or imagery cues to highlight these correspondences, such as making two representations visible and aligned spatially. The implementation will vary across disciplinary content due to the nature of disciplinary knowledge. In mathematics, for example, a teacher might explicitly highlight the relevance of multiple solution strategies or problems. The teacher might explicitly state a cue such as the following: "Solve this problem the same way as the last one". Alternatively, the teacher might use spatial cuing in which two problems are written on the board with their solution steps spatially aligned (see (Richland et al., 2007) for more examples from classroom practices). In science, this might include a teacher displaying a model and cuing the students into how to align this with the real-word scientific system (system of relationships). In history, this could look different, with cuing for example taking the form of asking students to develop an outline of key events or patterns of relationships such political systems and draw connections between similar processes across nation-states.
2. ***Build on relevant prior knowledge:*** Ensuring that reasoners have a strong knowledge base for conducting analogical reasoning is important, as this is a prerequisite for successful structure mapping and inferencing. Analogy may support building knowledge but relies upon adequate understanding of the representational structures to be used successfully. The timing may matter, such that opportunities for relational reasoning may support building schemas and more flexible transfer, but without adequate prior knowledge, students will need additional support to make these effective. Thus, we recommend that teachers check or enhance students' understanding of the key concepts they have acquired that will be needed to engage in relational reasoning when learning the new content. For example, when teaching proportions in a math class, before introducing proportions, teachers could begin with a few review questions of fraction, a key concept for learning proportions.
3. ***Support novices in a field in developing a relational mindset***: Expertise in many domains is evidenced through a focus on relations within the field (Chi et al., 1981; see Richland & Simms, 2015); thus, novices may build their expertise through learning that is organized relationally. Instructionally, this can be supported by using relational language and learning prompts that highlight the importance of retaining relational

versus object-based information or asking students to generate explanations that invite relational and causal thinking. For example, when teaching proportions, teachers could direct students' attention to the relation between lemon and water in two lemonades of different sizes and invite students to explain why the flavor of two lemonades stays the same despite different sizes. For young children, supporting domain-general opportunities to socialize attention to relations may be useful in fostering a relational mindset. That means children may bring those attentional patterns into new domain learning contexts, thereby supporting domain-specific knowledge building that is more expert-like and relational. This can be facilitated by socialization practices including asking open ended questions, using relational language, and setting up opportunities for children to notice relations or use relational language.

4. ***Use and Support Learners in Using Relational Language***: Using relational language can direct attention in a particular learning opportunity to key relational correspondences, or to a mode of thinking that invites relational reasoning. Relational language may include specific cues that highlight the relevant relation such as "above" or "halved." They may also be deictic terms that draw attention to correspondences across relational structures: "This one is like that one" (Alibali et al., 2014; Richland & Simms, 2015). More broadly, Higher-Order Thinking Talk (HOTT) can be used in many home or school contexts across tasks or domain contexts in order to shape children's relational skill development, their mindsets, and their school learning. These are relational words such as those listed above, as well as categorization, generalization, and related language processing. Most important is to set up a learning context to scaffold learners to use relational language themselves, which is more impactful than having the knowledgeable instructor to generate that language. Going back to the example of learning proportions, teachers could structure the lesson using comparison pedagogy, such as comparing different proportional problems with different cover stories and/or comparing different solution strategies for the same proportional problem. More importantly, teachers should have a review session intended for students to carry out the comparisons and their own explanations, which would elicit relational language and HOTT.

Altogether, these instructional practices are all feasible, low-resource strategies for encouraging learners to reason relationally while learning traditional academic subjects, leading to both transferable knowledge and a more expert-like approach to understanding domain knowledge. These practices encourage noticing the relevance of aligning and mapping relations, training students to look for these opportunities, or approaches to helping students who are by definition novices in a domain disattend to surface features and instead prioritize structural/more conceptual correspondences.

Author Contributions: Conceptualization, L.E.R. and H.Z.; methodology, L.E.R.; software, H.Z.; validation, L.E.R. and H.Z.; formal analysis, H.Z.; investigation, H.Z.; resources, L.E.R.; data curation, H.Z.; writing—original draft preparation, L.E.R. and H.Z.; writing—review and editing, L.E.R. and H.Z.; visualization, H.Z.; supervision, L.E.R.; project administration, H.Z.; funding acquisition, L.E.R. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Education Science, U.S. Department of Education (R305A240287-25), and the National Science Foundation (2141411 and 2327398). The views are of the authors and may not represent the funding agencies.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Topics	Included Articles
Significance of relational reasoning in the broad educational context	Darling-Hammond, L. (2022). Reimagining American education: Possible futures: The policy changes we need to get there. *Phi Delta Kappan*, *103*(8), 54–57. https://doi.org/10.1177/00317217221100012 Darling-Hammond, L., Alexander, M., & Hernández, L. E. (2024). *Redesigning high schools: 10 features for success.* Learning Policy Institute. National Research Council. (2012). *A framework for K–12 science education.* National Academies Press. NGSS Lead States. (2013). *Next generation science standards: For states, by states.* National Academies Press. OECD. (2025). *Trends shaping education 2025.* OECD Publishing. Richland, L. E., & Simms, N. (2015). Analogy, higher-order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 177–192. U.S. Department of Education, Office of Educational Technology. (2023). *Artificial intelligence and the future of teaching and learning*. World Economic Forum. (2023). *Defining education 4.0: A taxonomy for the future of learning*.
Theories of relational reasoning/higher-order thinking	Alexander, P. A., Dumas, D., Grossnickle, E. M., List, A., & Firetto, C. M. (2015). Measuring relational reasoning. *Journal of Experimental Education*, *84*, 119–151. Alexander, P. A., & The Disciplined Reading and Learning Research Laboratory. (2012). Reading into the future. *Educational Psychologist*, *47*(4), 259–280. Bloom, B. S. (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals; Handbook I: Cognitive Domain.* New York: Longman. Dumas, D., Alexander, P. A., & Grossnickle, E. M. (2013). Relational reasoning in educational contexts. *Educational Psychology Review*, *25*, 391–427. Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155–170. Gentner, D. (2003). Why we're so smart. In *Language in mind* (pp. 195–235). Gentner, D., Holyoak, K. J., & Kokinov, B. (Eds.). (2001). *The analogical mind.* MIT Press. Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity. *Cognitive Psychology*, *23*, 222–262.

Topics	Included Articles
Theories of relational reasoning/higher-order thinking	Holyoak, K.J., Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*, 295. Lewis, A., & Smith, D. (1993). Defining higher-order thinking. *Theory Into Practice*, *32*(3), 131–137. Resnick, L. B. (1987). *Education and learning to think.* National Academy Press. Spellman, B. A., & Holyoak, K. J. (1992). If Saddam is Hitler then who is George Bush? *Journal of Personality and Social Psychology*, *62*, 913–933.
Relational category	Doumas, L. A., & Hummel, J. E. (2013). Comparison and mapping facilitate relation discovery. *PLoS ONE*, *8*, e63889. Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge. *Psychological Bulletin, 142*, 729–757. Kalra, P. B., & Richland, L. E. (2022). Relational reasoning. *Mind, Brain, and Education*, *16*, 149–152.
Relational reasoning in problem-solving, creativity, and transfer	Catrambone, R., & Holyoak, K. J. (1989). Contextual limitations on transfer. *JEP:LMC*, *15*, 1147–1156. Dumas, D. (2018). Relational reasoning and divergent thinking. *Contemporary Educational Psychology*, *53*, 1–14. Dumas, D., Dong, Y., & Doherty, M. (2021). Creative expertise and analogical reasoning. *Mind, Brain, and Education*, *15*, 239–249. Dumas, D., Schmidt, L. C., & Alexander, P. A. (2016). Predicting creative problem solving. *Thinking Skills and Creativity*, *21*, 50–66. Dunbar, K. (1997). How scientists think. In *Conceptual structures and processes*. Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). Similarity in transfer. *Cognitive Psychology*, *25*, 524–575. Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355. Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38. Green, A. E. et al. (2012). Neural correlates of creativity in analogical reasoning. *JEP:LMC*, *38*, 264. Guilford, J. P. (1967b). *The nature of human intelligence*. McGraw-Hill.

Topics	Included Articles
Relational reasoning in problem-solving, creativity, and transfer	Halford, G. S., & Andrews, G. (2006). *Reasoning and problem solving.* Handbook of child psychology, Volume 2, Cognition, perception, and language. D. Kuhn, R.S. Siegler, W. Damon & R.M. Lerner (Eds). New Jersey USA: John Wiley & Sons, Inc. 557–608. https://doi.org/10.1002/9780470147658.chpsy0213 Kalra, P. B., & Richland, L. E. (2022). Relational reasoning: A foundation for higher cognition based on abstraction. *Mind, Brain, and Education*, *16*(2), 149–152. Loewenstein, J., & Gentner, D. (2005). Relational language. *Cognitive Psychology*, *50*, 315–353. Markman, A. B., & Wood, K. L. (Eds.). (2009). *Tools for innovation.* Oxford University Press. Novick, L. R. (1988). Analogical transfer and expertise. *JEP:LMC*, *14*, 510. Richland, L. E., & Simms, N. (2015). Analogy, higher-order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 177–192. Ross, B. H., & Kennedy, P. T. (1990). Generalizing from earlier examples. *JEP:LMC*, *16*, 42–55. Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*, 92–96. Skemp, R. R. (1979). *Intelligence, learning, and action.* Wiley. Vartanian, O. (2012). Neural systems for analogy and metaphor. *British Journal of Psychology*, *103*, 302–316.
Relational reasoning in domain-specific knowledge acquisition	Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, *48*, 87–113. Begolli, K. N., & Richland, L. E. (2016). Teaching mathematics by comparison: Analog visibility as a double-edged sword. *Journal of Educational Psychology*, *108*(2), 194–210. Borriello, G. A., Grenell, A., Vest, N. A., Moore, K., & Fyfe, E. R. (2023). Links between repeating and growing pattern knowledge and math outcomes in children and adults. *Child Development*, *94*(2), e103–e118. Braasch, J. L., & Goldman, S. R. (2010). The role of prior knowledge in learning from analogies in science texts. *Discourse Processes*, *47*(6), 447–479. Bustamante, A. S., Begolli, K. N., Alvarez-Vargas, D., Bailey, D. H., & Richland, L. E. (2022). Fraction ball: Playful and physically active fraction and decimal learning. *Journal of Educational Psychology*, *114*(6), 1307. Chan, J., & Schunn, C. (2015). The impact of analogies on creative concept generation: Lessons from an in vivo study in engineering design. *Cognitive Science*, *39*(1), 126–155. DeWolf, M., Bassok, M., & Holyoak, K. J. (2016). A set for relational reasoning: Facilitation of algebraic modeling by a fraction task. *Journal of Experimental Child Psychology*, *152*, 351–366.

Topics	Included Articles
	Dumas, D., Schmidt, L. C., & Alexander, P. A. (2016). Predicting creative problem solving in engineering design. *Thinking Skills and Creativity*, *21*, 50–66.
	Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg, & J. Davidson (Eds.), *The nature of insight* (pp. 365–396). MIT Press.
	Farrington-Flint, L., Canobi, K. H., Woor, C., & Faulkner, D. (2007). The role of relational reasoning in children's addition concepts. *British Journal of Developmental Psychology*, *25*, 227–246.
	Goswami, U., & Mead, F. (1992). Onset and rime awareness and analogies in reading. *Reading Research Quarterly*, *27*(2), 153–162.
	Halpern, D. F., Hansen, C., & Riefer, D. (1990). Analogies as an aid to understanding and memory. *Journal of Educational Psychology*, *82*(2), 298–306.
	Hansen, J., & Richland, L. E. (2020). Teaching and learning science through multiple representations: Intuitions and executive functions. *CBE—Life Sciences Education*, *19*(4), ar61.
	Klein, P. D., Piacente-Cimini, S., & Williams, L. A. (2007). The role of writing in learning from analogies. *Learning and Instruction*, *17*(6), 595–611.
Relational reasoning in domain-specific knowledge acquisition	Kok, E. M., de Bruin, A. B., Robben, S. G., & van Merriënboer, J. J. (2013). Learning radiological appearances of diseases: Does comparison help? *Learning and Instruction*, *23*, 90–97.
	Matlen, B. J., Gentner, D., & Franconeri, S. (2020). Spatial alignment facilitates visual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(5), 443–457.
	Mayer, R. E. (Ed.). (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge University Press.
	Murphy, P. K., Firetto, C. M., & Greene, J. A. (2017). Enriching students' scientific thinking through relational reasoning: Seeking evidence in texts, tasks, and talk. *Educational Psychology Review*, *29*, 105–117.
	Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 398–415.
	Richland, L. E., Holyoak, K. J., & Stigler, J. W. (2004). Analogy use in eighth-grade mathematics classrooms. *Cognition and Instruction*, *22*(1), 37–60.
	Spiro, R. J., Feltovich, R. L., & Anderson, D. K. (1988). *Multiple analogies for complex concepts: Antidotes for analogy-induced misconception in advanced knowledge acquisition*. Center for the Study of Reading Technical Report (No. 439). University of Illinois at Urbana–Champaign.
	White, C. S. (1998). The relationship between young children's analogical reasoning and mathematical learning. *Mathematical Cognition*, *4*(2), 103–123.

https://doi.org/10.3390/bs16020210

Topics	Included Articles
Relational reasoning in domain-specific knowledge acquisition	Zhao, H., Alexander, P. A., & Sun, Y. (2021). Relational reasoning's contributions to mathematical thinking and performance in Chinese elementary and middle-school students. *Journal of Educational Psychology*, *113*(2), 279–292. Zook, K. B., & Di Vesta, F. J. (1991). Effects of analogical processes on learning and misrepresentation. *Educational Psychology Review*, *3*(1), 41–72.
Relational reasoning in expertise development	Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–152. Goldwater, M. B., & Gentner, D. (2015). On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. *Cognition*, *137*, 137–153. Goldwater, M. B., Gentner, D., LaDue, N. D., & Libarkin, J. C. (2021). Analogy generation in science experts and novices. *Cognitive Science*, *45*(9), e13036. Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive Science*, *36*(5), 919–932.
Development of relational reasoning/higher-order thinking	Frausel, R. R., Richland, L. E., Levine, S. C., & Goldin-Meadow, S. (2021). Personal narrative as a "breeding ground" for higher-order thinking talk in early parent–child interactions. *Developmental Psychology*, *57*(4), 519–533. Frausel, R. R., Silvey, C., Freeman, C., Dowling, N., Richland, L. E., Levine, S. C., & Goldin-Meadow, S. (2020). The origins of higher-order thinking lie in children's spontaneous talk across the preschool years. *Cognition*, *200*, 104274. Frausel, R. R., Vollman, E., Muzard, A., Richland, L. E., Goldin-Meadow, S., & Levine, S. C. (2022). Developmental trajectories of early higher-order thinking talk differ for typically developing children and children with unilateral brain injuries. *Mind, Brain, and Education*, *16*(2), 153–166. Gentner, D. (1977). Children's performance on a spatial analogies task. *Child Development*, *48*, 1034–1039. Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, 47–59. Goswami, U. (2002). Inductive and deductive reasoning. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 282–302). Blackwell. Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*(6), 2797–2822. Krawczyk, D. C. (2012). The cognition and neuroscience of relational reasoning. *Brain Research*, *1428*, 13–23. Krawczyk, D. C., Hanten, G., Wilde, E. A., Li, X., Schnelle, K. P., Merkley, T. L., & Levin, H. S. (2010). Deficits in analogical reasoning in adolescents with traumatic brain injury. *Frontiers in Human Neuroscience*, *4*, 62.

Topics	Included Articles
Development of relational reasoning/higher-order thinking	Krawczyk, D. C., Morrison, R. G., Viskontas, I. V., Holyoak, K. J., Chow, T. W., Mendez, M. F., & Knowlton, B. J. (2008). Distraction during relational reasoning. *Neuropsychologia*, *46*(7), 2020–2032. Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning. *Journal of Cognitive Neuroscience*, *16*(2), 260–271. Pruden, S. M., Levine, S. C., & Huttenlocher, J. (2011). Children's spatial thinking. *Developmental Science*, *14*(6), 1417–1430. Rattermann, M. J., & Gentner, D. (1998). More evidence for a relational shift in the development of analogy. *Cognitive Development*, *13*(4), 453–478. Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning. *Journal of Experimental Child Psychology*, *94*(3), 249–273. Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities. *Science*, *323*(5916), 951–953. Silvey, C., Gentner, D., Richland, L. E., & Goldin-Meadow, S. (2023). Children's early spontaneous comparisons predict later analogical reasoning skills. *Open Mind*, *7*, 483–509. Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., & Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, *10*(2), 119–125.
Relational reasoning and explanation	Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples. *Review of Educational Research*, *70*, 181–214. Chi, M. T. H., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477. Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 260–276). Oxford University Press.
Relational reasoning and artificial intelligence	Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4176–4181. Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541.
Relational mindset	Goldwater, M. B., & Jamrozik, A. (2019). Can a relational mindset boost analogical retrieval? *Cognitive Research: Principles and Implications*, *4*, 1–16. Murphy, A. N., Zheng, Y., Shivaram, A., Vollman, E., & Richland, L. E. (2021). Bias and sensitivity to task constraints. *Journal of Experimental Child Psychology*, *202*, 104981.

Topics	Included Articles
Relational mindset	Simms, N. K., & Richland, L. E. (2019). Generating relations elicits a relational mindset in children. *Cognitive Science*, *43*(10), e12795. Vendetti, M. S., Matlen, B. J., Richland, L. E., & Bunge, S. A. (2015). Analogical reasoning in the classroom. *Mind, Brain, and Education*, *9*(2), 100–106. Vendetti, M. S., Wu, A., & Holyoak, K. J. (2014). Far-out thinking. *Psychological Science*, *25*(3), 1–6. Walker, C. M., Hubachek, S. Q., & Vendetti, M. S. (2018). Achieving abstraction. *Developmental Psychology*, *54*(10), 1833–1841. Zhao, H., Richland, L. E., Vollman, E., Lerner, B. S., Yeung, N. A., & Wong, J. (2025). Individual differences in attention to analogical relations. *Journal of Cognition and Development*. Zhao, H., Rose, E., & Richland, L. E. (2025). Priming multiplicative thinking across tasks. *SRCD Annual Conference*.
Analogy-based teaching practices	Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons. *Educational Psychologist*, *48*, 87–113. Alibali, M. W., Nathan, M. J., Wolfgram, M. S., Church, R. B., Jacobs, S. A., Johnson Martinez, C., & Knuth, E. J. (2014). How teachers link ideas in mathematics instruction. *Cognition and Instruction*, *32*(1), 65–100. Begolli, K. N., Richland, L. E., Jaeggi, S. M., Lyons, E. M., Klostermann, E. C., & Matlen, B. J. (2018). Executive function in learning mathematics by comparison. *Thinking & Reasoning*, *24*(2), 280–313. Blessing, S. B., & Ross, B. H. (1996). The effect of problem content on categorization and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 792–810. Durkin, K., Rittle-Johnson, B., Star, J. R., & Loehr, A. (2023). Comparing and discussing multiple strategies. *Journal of Experimental Education*, *91*(1), 1–19. Richland, L. E., & McDonough, I. M. (2010). Learning by analogy. *Contemporary Educational Psychology*, *35*(1), 28–43. Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science*, *316*(5828), 1128–1129. Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate knowledge? *Journal of Educational Psychology*, *99*(3), 561–574. Rittle-Johnson, B., Star, J. R., & Durkin, K. (2009). The importance of prior knowledge when comparing examples. *Journal of Educational Psychology*, *101*(4), 836–852. Rittle-Johnson, B., Star, J. R., & Durkin, K. (2017). The power of comparison in mathematics instruction. In *Mathematical cognition and learning* (Vol. 3). Elsevier.

Topics	Included Articles
	Star, J. R., & Rittle-Johnson, B. (2008). Flexibility in problem solving. *Learning and Instruction*, *18*(6), 565–579.
Analogy-based teaching practices	Star, J. R., Pollack, C., Durkin, K., Rittle-Johnson, B., Lynch, K., Newton, K., & Gogolen, C. (2015). Learning from comparison in algebra. *Contemporary Educational Psychology*, *40*, 41–54.
	Williams, T., Ferraro, D., Roey, S., Brenwald, S., Kastberg, D., Jocelyn, L., & Stearns, P. (2009). *TIMSS 2007 U.S. technical report and user guide.* NCES.

Notes

[1] The studies reviewed in this article were selected for their relevance to the major themes of the manuscript and the key theories in the field of relational reasoning, as well as its subfields. These subfields include relational category acquisition, the role of relational reasoning in other domain-general skills (e.g., problem solving, knowledge transfer) and domain-specific knowledge (e.g., mathematics), spontaneous relational attention, and evidence-based teaching practices that support relational reasoning. We selected classic theoretical work, such as Dedre Gentner's structure mapping theory, foundational empirical studies, and more recent empirical work on spontaneous relational attention and large language models. The core set of articles was identified primarily through the authors' expertise in the field and hand-searching publications by prominent researchers in the field. This set was then expanded through snowballing methods by reviewing the reference lists of these key articles. Please see a complete list of studies included in this review in the Appendix A.

References

Alexander, P. A., Dumas, D., Grossnickle, E. M., List, A., & Firetto, C. M. (2015). Measuring relational reasoning. *Journal of Experimental Education*, *84*(1), 119–151. [CrossRef]

Alexander, P. A., & The Disciplined Reading and Learning Research Laboratory. (2012). Reading into the future: Competence for the 21st century. *Educational Psychologist*, *47*(4), 259–280. [CrossRef]

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, *48*, 87–113. [CrossRef]

Alibali, M. W., Nathan, M. J., Wolfgram, M. S., Church, R. B., Jacobs, S. A., Johnson Martinez, C., & Knuth, E. J. (2014). How teachers link ideas in mathematics instruction using speech and gesture: A corpus analysis. *Cognition and Instruction*, *32*(1), 65–100. [CrossRef]

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives* (complete ed.). Longman.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, *70*, 181–214. [CrossRef]

Begolli, K. N., & Richland, L. E. (2016). Teaching mathematics by comparison: Analog visibility as a double-edged sword. *Journal of Educational Psychology*, *108*(2), 194–210. [CrossRef]

Begolli, K. N., Richland, L. E., Jaeggi, S. M., Lyons, E. M., Klostermann, E. C., & Matlen, B. J. (2018). Executive function in learning mathematics by comparison: Incorporating everyday classrooms into the science of learning. *Thinking & Reasoning*, *24*(2), 280–313. [CrossRef]

Blessing, S. B., & Ross, B. H. (1996). The effect of problem content on categorization and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 792–810. [CrossRef]

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals; handbook I: Cognitive domain*. Longman.

Borriello, G. A., Grenell, A., Vest, N. A., Moore, K., & Fyfe, E. R. (2023). Links between repeating and growing pattern knowledge and math outcomes in children and adults. *Child Development*, *94*(2), e103–e118. [CrossRef]

Braasch, J. L., & Goldman, S. R. (2010). The role of prior knowledge in learning from analogies in science texts. *Discourse Processes*, *47*(6), 447–479. [CrossRef]

Bustamante, A. S., Begolli, K. N., Alvarez-Vargas, D., Bailey, D. H., & Richland, L. E. (2022). Fraction ball: Playful and physically active fraction and decimal learning. *Journal of Educational Psychology*, *114*(6), 1307. [CrossRef]

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1147–1156. [CrossRef]

Chan, J., & Schunn, C. (2015). The impact of analogies on creative concept generation: Lessons from an in vivo study in engineering design. *Cognitive Science*, *39*(1), 126–155. [CrossRef] [PubMed]

Chi, M. T. H., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477. [CrossRef]

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*, 121–152. [CrossRef]

Darling-Hammond, L. (2022). Reimagining American education: Possible futures: The policy changes we need to get there. *Phi Delta Kappan*, *103*(8), 54–57. [CrossRef]

Darling-Hammond, L., Alexander, M., & Hernández, L. E. (2024). *Redesigning high schools: 10 features for success*. Learning Policy Institute. [CrossRef]

DeWolf, M., Bassok, M., & Holyoak, K. J. (2016). A set for relational reasoning: Facilitation of algebraic modeling by a fraction task. *Journal of Experimental Child Psychology*, *152*, 351–366. [CrossRef]

Doumas, L. A., & Hummel, J. E. (2013). Comparison and mapping facilitate relation discovery and predication. *PLoS ONE*, *8*(6), e63889. [CrossRef]

Dumas, D. (2018). Relational reasoning and divergent thinking: An examination of the threshold hypothesis with quantile regression. *Contemporary Educational Psychology*, *53*, 1–14. [CrossRef]

Dumas, D., Alexander, P. A., & Grossnickle, E. M. (2013). Relational reasoning and its manifestations in the educational context: A systematic review of the literature. *Educational Psychology Review*, *25*, 391–427. [CrossRef]

Dumas, D., Dong, Y., & Doherty, M. (2021). The influence of creative expertise on the sensitivity and selectivity of analogical reasoning. *Mind, Brain, and Education*, *15*(3), 239–249. [CrossRef]

Dumas, D., Schmidt, L. C., & Alexander, P. A. (2016). Predicting creative problem solving in engineering design. *Thinking Skills and Creativity*, *21*, 50–66. [CrossRef]

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg, & J. Davidson (Eds.), *The nature of insight* (pp. 365–396). MIT Press.

Dunbar, K. (1997). How scientists think: On-line creativity and conceptual change in science. In T. B. Ward, S. M. Smith, & S. Vaid (Eds.), *Conceptual structures and processes: Emergence, discovery and change* (pp. 461–493). American Psychological Association.

Durkin, K., Rittle-Johnson, B., Star, J. R., & Loehr, A. (2023). Comparing and discussing multiple strategies: An approach to improving algebra instruction. *Journal of Experimental Education*, *91*(1), 1–19. [CrossRef]

Farrington-Flint, L., Canobi, K. H., Woor, C., & Faulkner, D. (2007). The role of relational reasoning in children's addition concepts. *British Journal of Developmental Psychology*, *25*, 227–246. [CrossRef]

Frausel, R. R., Richland, L. E., Levine, S. C., & Goldin-Meadow, S. (2021). Personal narrative as a "breeding ground" for higher-order thinking talk in early parent–child interactions. *Developmental Psychology*, *57*(4), 519. [CrossRef]

Frausel, R. R., Silvey, C., Freeman, C., Dowling, N., Richland, L. E., Levine, S. C., Raudenbush, S., & Goldin-Meadow, S. (2020). The origins of higher-order thinking lie in children's spontaneous talk across the pre-school years. *Cognition*, *200*, 104274. [CrossRef]

Frausel, R. R., Vollman, E., Muzard, A., Richland, L. E., Goldin-Meadow, S., & Levine, S. C. (2022). Developmental trajectories of early higher-order thinking talk differ for typically developing children and children with unilateral brain injuries. *Mind, Brain, and Education*, *16*(2), 153–166. [CrossRef]

Gentner, D. (1977a). Children's performance on a spatial analogies task. *Child Development*, *48*, 1034–1039. [CrossRef]

Gentner, D. (1977b). If a tree had a knee, where would it be? Children's Performance on simple spatial metaphors. *Papers and Reports on Child Language Development*, *13*, 157–164.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170. [CrossRef]

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, *59*(1), 47–59. [CrossRef]

Gentner, D. (2003). Why we're so smart. In *Language in mind: Advances in the study of language and thought* (pp. 195–235). Boston Review.

Gentner, D., Holyoak, K. J., & Kokinov, B. (Eds.). (2001). *The analogical mind: Perspectives from cognitive science*. MIT Press.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology*, *25*, 524–575. [CrossRef]

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355. [CrossRef]

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38. [CrossRef]

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, *23*(2), 222–262. [CrossRef] [PubMed]

Goldwater, M. B., & Gentner, D. (2015). On the acquisition of abstract knowledge: Structural alignment and explication in learning causal system categories. *Cognition*, *137*, 137–153. [CrossRef]

Goldwater, M. B., Gentner, D., LaDue, N. D., & Libarkin, J. C. (2021). Analogy generation in science experts and novices. *Cognitive Science*, *45*(9), e13036. [CrossRef]

Goldwater, M. B., & Jamrozik, A. (2019). Can a relational mindset boost analogical retrieval? *Cognitive Research: Principles and Implications*, *4*, 47. [CrossRef] [PubMed]

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*(7), 729–757. [CrossRef]

Goswami, U. (2002). Inductive and deductive reasoning. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 282–302). Blackwell Publishing. [CrossRef]

Goswami, U., & Mead, F. (1992). Onset and rime awareness and analogies in reading. *Reading Research Quarterly*, *27*(2), 153–162. [CrossRef]

Green, A. E., Kraemer, D. J., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2012). Neural correlates of creativity in analogical reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 264. [CrossRef]

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Halford, G. S., & Andrews, G. (2006). Reasoning and problem solving. In D. Kuhn, & R. Siegler (Eds.), *Handbook of child psychology: Cognition, perception, and language* (6th ed., Vol. 2, pp. 557–608). John Wiley & Sons.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*, 803–865. [CrossRef] [PubMed]

Halpern, D. F., Hansen, C., & Riefer, D. (1990). Analogies as an aid to understanding and memory. *Journal of Educational Psychology*, *82*(2), 298. [CrossRef]

Hansen, J., & Richland, L. E. (2020). Teaching and learning science through multiple representations: Intuitions and executive functions. *CBE—Life Sciences Education*, *19*(4), ar61. [CrossRef]

Holyoak, K. J., & Thagard, P. (1989). A computational model of analogical problem solving. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 242–266). Cambridge University Press.

Kalra, P. B., & Richland, L. E. (2022). Relational reasoning: A foundation for higher cognition based on abstraction. *Mind, Brain, and Education*, *16*(2), 149–152. [CrossRef]

Klein, P. D., Piacente-Cimini, S., & Williams, L. A. (2007). The role of writing in learning from analogies. *Learning and Instruction*, *17*(6), 595–611. [CrossRef]

Kok, E. M., de Bruin, A. B., Robben, S. G., & van Merriënboer, J. J. (2013). Learning radiological appearances of diseases: Does comparison help? *Learning and Instruction*, *23*, 90–97. [CrossRef]

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*(6), 2797–2822. [CrossRef]

Krawczyk, D. C. (2012). The cognition and neuroscience of relational reasoning. *Brain Research*, *1428*, 13–23. [CrossRef]

Krawczyk, D. C., Hanten, G., Wilde, E. A., Li, X., Schnelle, K. P., Merkley, T. L., Vasquez, A. C., Cook, L. G., McClelland, M., Chapman, S. B., & Levin, H. S. (2010). Deficits in analogical reasoning in adolescents with traumatic brain injury. *Frontiers in Human Neuroscience*, *4*, 62. [CrossRef] [PubMed]

Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., Miller, B. L., & Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia*, *46*(7), 2020–2032. [CrossRef]

Lewis, A., & Smith, D. (1993). Defining higher order thinking. *Theory into Practice*, *32*(3), 131–137. [CrossRef]

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, *50*(4), 315–353. [CrossRef] [PubMed]

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. [CrossRef]

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak, & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 260–276). Oxford University Press.

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4176–4181. [CrossRef] [PubMed]

Markman, A. B., & Wood, K. L. (Eds.). (2009). *Tools for innovation: The science behind the practical methods that drive new ideas*. Oxford University Press.

Matlen, B., Gentner, D., & Franconeri, S. (2020). Spatial alignment facilitates visual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(5), 443–457. [CrossRef]

Mayer, R. E. (Ed.). (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge University Press. [CrossRef]

Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, *16*(2), 260–271. [CrossRef]

Murphy, A. N., Zheng, Y., Shivaram, A., Vollman, E., & Richland, L. E. (2021). Bias and sensitivity to task constraints in spontaneous relational attention. *Journal of Experimental Child Psychology*, *202*, 104981. [CrossRef]

Murphy, P. K., Firetto, C. M., & Greene, J. A. (2017). Enriching students' scientific thinking through relational reasoning: Seeking evidence in texts, tasks, and talk. *Educational Psychology Review*, *29*, 105–117. [CrossRef]

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academies Press.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 510. [CrossRef]

Novick, L. R., & Holyoak, K. J. (1991). Mathematical problem solving by analogy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 398–415. [CrossRef]

OECD. (2025). *Trends shaping education 2025*. OECD Publishing. [CrossRef]

Pruden, S. M., Levine, S. C., & Huttenlocher, J. (2011). Children's spatial thinking: Does talk about the spatial world matter? *Developmental Science*, *14*(6), 1417–1430. [CrossRef] [PubMed]

Rattermann, M. J., & Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task. *Cognitive Development*, *13*(4), 453–478. [CrossRef]

Resnick, L. B. (1987). *Education and learning to think*. National Academy Press.

Richland, L. E., Holyoak, K. J., & Stigler, J. W. (2004). Analogy use in eighth-grade mathematics classrooms. *Cognition and Instruction*, *22*(1), 37–60. [CrossRef]

Richland, L. E., & McDonough, I. M. (2010). Learning by analogy: Discriminating between potential analogs. *Contemporary Educational Psychology*, *35*(1), 28–43. [CrossRef]

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94*(3), 249–273. [CrossRef]

Richland, L. E., & Simms, N. (2015). Analogy, higher-order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 177–192. [CrossRef]

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science*, *316*(5828), 1128–1129. [CrossRef]

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, *99*(3), 561–574. [CrossRef]

Rittle-Johnson, B., Star, J. R., & Durkin, K. (2009). The importance of prior knowledge when comparing examples: Influences on conceptual and procedural knowledge of equation solving. *Journal of Educational Psychology*, *101*(4), 836–852. [CrossRef]

Rittle-Johnson, B., Star, J. R., & Durkin, K. (2017). The power of comparison in mathematics instruction: Experimental evidence from classrooms. In D. C. Geary, D. B. Berch, & K. M. Koepke (Eds.), *Mathematical cognition and learning* (Vol. 3, pp. 273–296). Elsevier.

Ross, B. H., & Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 42–55. [CrossRef]

Rottman, B. M., Gentner, D., & Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. *Cognitive Science*, *36*(5), 919–932. [CrossRef] [PubMed]

Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, *323*(5916), 951–953. [CrossRef]

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*(1), 92–96. [CrossRef]

Silvey, C., Gentner, D., Richland, L. E., & Goldin-Meadow, S. (2023). Children's early spontaneous comparisons predict later analogical reasoning skills: An investigation of parental influence. *Open Mind*, *7*, 483–509. [CrossRef]

Simms, N. K., & Richland, L. E. (2019). Generating relations elicits a relational mindset in children. *Cognitive Science*, *43*(10), e12795. [CrossRef]

Skemp, R. R. (1979). *Intelligence, learning, and action: A foundation for theory and practice in education*. Wiley.

Spellman, B. A., & Holyoak, K. J. (1992). If Saddam is Hitler then who is George Bush? Analogical mapping between systems of social roles. *Journal of Personality and Social Psychology*, *62*(6), 913–933. [CrossRef]

Spiro, R. J., Feltovich, R. L., & Anderson, D. K. (1988). *Multiple analogies for complex concepts: Antidotes for analogy-induced misconception in advanced knowledge acquisition* (No. 439). University of Illinois at Urbana-Champaign.

Star, J. R., Pollack, C., Durkin, K., Rittle-Johnson, B., Lynch, K., Newton, K., & Gogolen, C. (2015). Learning from comparison in algebra. *Contemporary Educational Psychology*, *40*, 41–54. [CrossRef]

Star, J. R., & Rittle-Johnson, B. (2008). Flexibility in problem solving: The case of equation solving. *Learning and Instruction*, *18*(6), 565–579. [CrossRef]

U.S. Department of Education, Office of Educational Technology. (2023). *Artificial intelligence and future of teaching and learning: Insights and recommendations*. U.S. Department of Education.

Vartanian, O. (2012). Dissociable neural systems for analogy and metaphor: Implications for the neuroscience of creativity. *British Journal of Psychology*, *103*(3), 302–316. [CrossRef]

Vendetti, M. S., Matlen, B. J., Richland, L. E., & Bunge, S. A. (2015). Analogical reasoning in the classroom: Insights from cognitive science. *Mind, Brain, and Education*, *9*(2), 100–106. [CrossRef]

Vendetti, M. S., Wu, A., & Holyoak, K. J. (2014). Far-out thinking: Generating solutions to distant analogies promotes relational thinking. *Psychological Science*, *25*(4), 928–933. [CrossRef]

Walker, C. M., Hubachek, S. Q., & Vendetti, M. S. (2018). Achieving abstraction: Generating far analogies promotes relational reasoning in children. *Developmental Psychology*, *54*(10), 1833–1841. [CrossRef] [PubMed]

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menezes Santos, M., Thomas, C. R., & Miller, B. L. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, *10*(2), 119–125. [CrossRef]

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526–1541. [CrossRef]

White, C. S. (1998). The relationship between young children's analogical reasoning and mathematical learning. *Mathematical Cognition*, *4*(2), 103–123. [CrossRef]

Williams, T., Ferraro, D., Roey, S., Brenwald, S., Kastberg, D., Jocelyn, L., & Stearns, P. (2009). *TIMSS 2007 US technical report and user guide*. National Center for Education Statistics, Institute of Education Sciences, US Department of Education.

World Economic Forum. (2023). *Defining education 4.0: A taxonomy for the future of learning*. Available online: https://www3.weforum.org/docs/WEF_Defining_Education_4.0_2023.pdf (accessed on 18 December 2025).

Zhao, H., Alexander, P. A., & Sun, Y. (2021). Relational reasoning's contributions to mathematical thinking and performance in Chinese elementary and middle-school students. *Journal of Educational Psychology*, *113*(2), 279–292. [CrossRef]

Zhao, H., Richland, L. E., Vollman, E., Lerner, B. S., Yeung, N. A., & Wong, J. (2025a). Individual differences in attention to analogical relations. *Journal of Cognition and Development*, *26*(3), 464–492. [CrossRef]

Zhao, H., Rose, E., & Richland, L. E. (2025b, May 1–3). *Priming multiplicative thinking across tasks and its implications for learning: Evidence from an intervention study* [Poster presentation]. Society for Research in Child Development Annual Conference, Minneapolis, MN, USA.

Zook, K. B., & Di Vesta, F. J. (1991). Effects of analogical processes on learning and misrepresentation. *Educational Psychology Review*, *3*(1), 41–72. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Review

Working Memory and Instructional Fit: Reintroducing Aptitude–Treatment Interaction in Education Research

Faria Sana [1,*] and Barbara Fenesi [2]

[1] Faculty of Humanities and Social Sciences, Athabasca University, Athabasca, AB T9S 3A3, Canada
[2] Faculty of Education, Western University, London, ON N6G 1G7, Canada; bfenesi@uwo.ca
* Correspondence: fsana@athabascau.ca

Abstract

Working memory (WM) is a cognitive system with limited capacity that enables individuals to focus on goal-relevant information while filtering distractions and integrating new knowledge. Differences in WM capacity influence how students engage with instructional materials, shaping the effectiveness of educational interventions. This raises the following question: which instructional methods work best, for whom, and under what conditions? The aptitude–treatment interaction (ATI) framework addresses this by examining how learning outcomes are influenced by the alignment between cognitive abilities, such as WM, and instructional strategies. This paper reviews WM's role in learning and academic achievement, explores ATI's relevance in contemporary education research, and highlights how this framework can guide the development of adaptive instructional strategies that support diverse learners.

Keywords: working memory; academic achievement; aptitude–treatment interaction

1. Introduction

Over the past century, cognitive psychology has significantly advanced our understanding of human learning and memory. Research highlights that learning outcomes are influenced by underlying cognitive abilities, which vary meaningfully among individuals (Rohrer & Pashler, 2010; Dunlosky & Rawson, 2019). Contemporary educational research increasingly acknowledges that the effectiveness of instructional techniques cannot be universally assumed but depends on individual learner characteristics. However, despite broad recognition, cognitive aptitudes such as working memory (WM) remain underutilized in instructional design and practical educational contexts. These abilities are often treated as random variance rather than as key determinants that explain differential instructional effectiveness (Sala & Gobet, 2020). To address this gap, this paper focuses on working memory capacity (WMC) within the aptitude–treatment interaction (ATI) framework, highlighting how aligning instructional strategies with learners' WMC can significantly enhance learning outcomes.

The ATI framework, originally conceptualized by Cronbach and Snow (1977), seeks to identify instructional treatments whose effectiveness varies systematically based on learner aptitudes—relatively stable cognitive, affective, or motivational characteristics (Preacher & Sterba, 2018; Rohrer & Pashler, 2010). Instructional treatments refer to pedagogical strategies designed to facilitate learning, such as multimedia instruction, retrieval practice, spaced repetition, and feedback provision. An aptitude–treatment interaction occurs when instructional outcomes differ systematically based on learners' cognitive characteristics (Mayer, 2017).

The ATI perspective is grounded in the assumption that learners with different cognitive abilities engage in qualitatively distinct cognitive processes rather than simply learning at different rates (Rohrer & Pashler, 2010). Instructional strategies beneficial for one group, such as individuals with high WMC, may provide limited or even negative outcomes for another group, such as individuals with low WMC. Empirical research supports this assumption, demonstrating that some instructional methods designed to optimize WM resources disproportionately benefit learners with lower WMC without negatively impacting higher-capacity peers (Fyfe & Rittle-Johnson, 2017; Peng & Fuchs, 2016).

Although numerous cognitive traits contribute to learning—such as executive function, attentional control, and fluid intelligence—this paper focuses specifically on WMC for both theoretical and practical reasons. First, WMC is a well-defined, reliably measured construct that plays a central role in managing the encoding, maintenance, and manipulation of information during learning tasks. It is strongly predictive of academic achievement across subject areas, even when controlling for broader intelligence measures (Alloway & Alloway, 2010). Second, WMC supports and overlaps with many other aptitudes, serving as a cognitive bottleneck that affects performance in tasks requiring attention, reasoning, or self-regulation. As such, it offers a conceptually coherent and empirically tractable target for investigating ATI. Finally, unlike broader constructs such as intelligence or executive function, WMC is more easily aligned with specific instructional demands, making it especially well suited for designing personalized, evidence-based educational interventions.

To this end, our paper aims to achieve three primary objectives: (1) to provide a comprehensive theoretical and empirical overview of WM as a cognitive aptitude and its influence on educational performance, (2) to briefly review recent ATI research that examines the interactions between WMC and instructional strategies, and (3) to discuss the key methodological considerations while outlining both the practical and theoretical implications for educators and researchers who seek to apply the ATI framework in educational settings. The structure of the paper follows these objectives.

2. WM and Cognitive Performance

The definition, structure, and function of WM have been extensively debated. However, three fundamental aspects are well established in the literature. First, WM is closely linked to attentional control and is not merely a passive storage system. Rather, it plays an active role in keeping task-relevant information accessible while filtering out distractions (Oberauer, 2019; Shipstead et al., 2016). Second, WMC is a relatively stable cognitive trait, meaning that although it can fluctuate depending on factors such as fatigue or motivation, it remains constrained by biological limits and does not change significantly over time (Redick et al., 2016; Xu et al., 2018). Third, deficits in WM significantly impact daily functioning and academic performance, and they are commonly associated with neurodevelopmental disorders such as ADHD and dyslexia (Snyder et al., 2015).

For educational contexts, WM is best understood as comprising two interrelated components: domain-general control processes and domain-specific storage mechanisms. The domain-general aspect of WM includes attentional control and strategic retrieval from long-term memory, which help learners maintain focus on relevant information while suppressing distractions (Shipstead et al., 2016; Unsworth & Robison, 2020). The domain-specific aspect refers to the short-term storage of different types of information, including phonological and auditory input, which are processed in the verbal store, and visual, spatial, and movement-related information, which are handled by the visuospatial store (Logie et al., 2021). These two subsystems work in tandem, linking new information

with pre-existing knowledge stored in long-term memory, while domain-general processes ensure that these connections remain active (Cowan, 2017; Logie et al., 2021).

WMC is typically measured using complex span tasks that assess both storage and processing simultaneously. A widely used example is the reading span task, in which participants read sentences while recalling the final word of each. Another common measure is the operation span task, where individuals solve math problems while remembering numerical information. The symmetry span task, which requires recalling visual patterns while making spatial judgments, is also frequently employed (Unsworth et al., 2005; Foster et al., 2015). Despite differences in format, all these tasks assess the same fundamental ability: the capacity to juggle multiple pieces of information while engaging in higher-order cognitive processes. These tasks reliably differentiate between high- and low-WMC individuals, and refinements in measurement techniques have further strengthened their predictive power for academic performance (Draheim et al., 2016; Redick et al., 2016).

Individual differences in WMC are strongly associated with higher-order thinking skills, including reasoning, comprehension, and problem-solving. Reading comprehension, for example, relies on the ability to hold onto words and phrases long enough to integrate meaning across sentences, resolve ambiguities, and draw inferences. Research has consistently shown that WMC is a strong predictor of reading ability, independent of general intelligence and verbal proficiency (Peng et al., 2018). Similarly, in mathematics, WM is essential for multi-step problem-solving, where students must retain intermediate calculations while applying formulas or reasoning strategies (Friso-van den Bos et al., 2015).

Beyond complex academic tasks, WM also plays a fundamental role in basic cognitive functions such as attention control, inhibition, and interference management. Individuals with higher WMC are better at maintaining focus on relevant information while ignoring distractions, which explains why they tend to perform better in cognitively demanding environments (Unsworth & Robison, 2020). In contrast, individuals with lower WMC struggle more with sustained attention and are more susceptible to mind-wandering (Shipstead et al., 2016). WM is also strongly linked to proactive interference, a phenomenon in which previously learned information disrupts the retrieval of new material. Studies indicate that lower-WMC individuals generate retrieval cues that are less distinct, making it more difficult to access relevant information when needed (Unsworth & Engle, 2007; Unsworth et al., 2012b).

One of the most widely studied relationships in WM research is its connection to fluid intelligence (gF)—the capacity to reason abstractly, identify patterns, and solve novel problems (Burgess et al., 2011). Numerous studies have demonstrated a strong correlation between WMC and gF, with neuroimaging evidence indicating that both rely on overlapping brain networks in the frontal and parietal regions, which are crucial for executive function and controlled processing (Burgess et al., 2011; Finn et al., 2015). Individuals with higher WMC tend to exhibit greater neural efficiency, meaning they can perform cognitive tasks with reduced mental effort. This efficiency is particularly evident in tasks that require filtering out irrelevant information or maintaining multiple sources of information simultaneously (Shipstead et al., 2014).

The link between WM, attentional control, and intelligence has significant implications for learning. When instructional materials present high cognitive demands, learners with lower WMC may be more susceptible to cognitive overload, which occurs when task complexity exceeds an individual's processing capacity. This is particularly problematic in disciplines such as mathematics and science, where students must integrate multiple sources of information simultaneously. If instructional strategies do not account for these limitations, lower-WMC students may struggle to process content effectively, even if they possess sufficient domain knowledge.

In summary, WM serves as a fundamental component of learning, problem-solving, and daily cognitive functioning. Its role in managing attention, suppressing distractions, and integrating new information makes it a critical predictor of academic success (Alloway & Alloway, 2010). Given the broad range of cognitive processes affected by WMC, understanding how instructional strategies interact with WM differences is essential for designing effective educational interventions.

3. WM and Academic Achievement

The previous section established the essential role of WM in cognition and learning. This section examines how individual differences in WMC predict academic success across core educational domains, including reading comprehension, writing, mathematics, second language acquisition, and science learning. In addition to subject-specific outcomes, WMC also influences fundamental classroom behaviors such as note-taking, following instructions, and managing distractions. Synthesizing recent research on these relationships highlights the broad impact of WM on both academic achievement and classroom engagement.

3.1. Role of WM in Academic Domains

3.1.1. Reading Comprehension

WMC is widely recognized as a key determinant of reading comprehension. Successful reading requires the continuous processing, integration, and updating of textual information. Individuals with higher WMC manage complex and lengthy passages more effectively, particularly when dealing with embedded clauses, ambiguous pronouns, or abstract concepts (Peng et al., 2018; Swanson et al., 2016). Even when controlling for factors such as verbal ability, decoding skills, and attention, WMC remains a strong predictor of reading comprehension outcomes (Kim, 2016). This relationship holds across diverse languages and writing systems, suggesting that WM plays a universal role in reading development. In contrast, lower-WMC readers are more likely to experience comprehension breakdowns, particularly when required to draw inferences or integrate information across multiple sentences. These findings highlight the importance of instructional strategies that support lower-WMC learners by reducing cognitive load in reading tasks.

3.1.2. Written Expression

Writing places substantial demands on WM, as it requires the simultaneous execution of multiple cognitive processes. Writers must generate ideas, structure sentences, maintain coherence, and monitor for grammatical accuracy while holding key elements in WM. Research has shown that WMC is a significant predictor of writing proficiency, particularly in areas such as syntactic complexity, coherence, and mechanical accuracy (Olive, 2014; Vanderberg & Swanson, 2007). Lower-WMC learners often struggle to maintain cohesion in extended writing tasks and are more prone to errors, particularly under time constraints or increased cognitive load. Studies on writing development suggest that transcription fluency and self-regulation are key mediators in how cognitive resources impact writing quality (Limpo & Alves, 2013). These findings reinforce the role of WMC in written expression and highlight the importance of instructional strategies that alleviate cognitive strain during writing tasks.

3.1.3. Second Language Acquisition

Second language acquisition also relies heavily on WMC, particularly in vocabulary learning, syntactic processing, and phonological representation. Research has demonstrated that phonological short-term memory is crucial for acquiring new vocabulary, as higher-

WMC learners retain and integrate novel linguistic patterns more effectively. For instance, Baddeley (2003) discusses the role of WM in language learning, highlighting its importance in vocabulary acquisition. Additionally, studies indicate that high-WMC individuals manage syntactic complexity more efficiently and adapt to new language structures with greater ease across different language pairs. Juffs and Harrington (2011) review the role of WM in second language learning and processing, emphasizing its impact on syntactic processing. Similarly, Linck et al. (2014) found that WMC is a significant predictor of language learning success, particularly in managing complex syntactic structures. In contrast, lower-WMC learners require more repetition and structured exposure to achieve comparable proficiency levels. These findings underscore the importance of instructional strategies that reduce cognitive load during second language learning to support learners with lower WMC.

3.1.4. Mathematics and Problem-Solving

Mathematics is another domain where WM plays a central role, particularly in multi-step calculations, algebraic reasoning, and problem-solving. Meta-analyses confirm that WMC is one of the strongest cognitive predictors of mathematical proficiency, particularly in tasks requiring sustained attention, information manipulation, and mental arithmetic (Peng et al., 2016; Zhang et al., 2023). Students with higher WMC are more adept at retaining numerical values, tracking intermediate steps, and applying problem-solving strategies efficiently. Conversely, those with mathematical learning difficulties frequently exhibit lower WMC, which affects their ability to retrieve math facts, maintain WM representations, and process multi-step problems (Geary et al., 2017; Fung & Swanson, 2017). These effects are especially pronounced in abstract problem-solving tasks that require learners to juggle multiple pieces of information without external aids.

3.1.5. Science Learning and Reasoning

Science learning often involves complex cognitive tasks such as reasoning through causal relationships, analyzing experimental data, and interpreting scientific texts. These tasks require the integration of multiple information sources simultaneously, including textual explanations, diagrams, and data representations. Research indicates that students with higher WMC may perform better in problem-solving environments, as they can manage and coordinate multiple information streams more effectively. For instance, Wiley and Jarosz (2012) discuss the relationship between WMC and problem-solving abilities, highlighting that higher WMC can be beneficial in analytical problem-solving contexts. Similarly, Tolmie et al. (2016) explore the role of executive functions, including WM, in children's science learning, suggesting that these cognitive skills are foundational for developing scientific reasoning abilities. Conversely, students with lower WMC may struggle to synthesize information from various sources, making it more challenging to grasp complex scientific concepts. As science education increasingly incorporates multimodal instruction, understanding how WMC interacts with different instructional formats remains an important area for future research.

While WMC influences a wide range of learning outcomes, this review focuses specifically on academic achievement due to its relevance for instructional design and evidence-based intervention. Importantly, WMC does not operate uniformly across subjects; each domain imposes distinct cognitive demands, which in turn affect how ATI-informed strategies should be applied. For example, lower-WMC learners may benefit more from scaffolding and worked examples in mathematics, while segmentation and dual-channel multimedia may be better suited to support science learning. In contrast, reading comprehension and second language acquisition often place greater strain on phonological

memory, requiring repetition and structured exposure. These differences underscore the importance of tailoring instructional supports not only to cognitive aptitude but also to the demands of the academic subject.

3.2. Role of WM in Classroom Skills

Beyond subject-specific outcomes, WMC also affects essential classroom behaviors, including reasoning, participation, note-taking, and behavior regulation. While these skills are often overlooked in traditional assessments of academic achievement, they significantly influence learning outcomes and engagement.

3.2.1. Classroom Reasoning and Participation

Inferential reasoning and logical deduction are essential components of class discussions and problem-solving activities, relying on the capacity to hold and manipulate multiple pieces of information simultaneously—a function closely tied to WM (Kyllonen & Christal, 1990). Research indicates that students with higher WMC often excel in tasks requiring the integration of prior knowledge with new information, facilitating more active engagement in complex cognitive activities (Alloway & Copello, 2013). In contrast, students with lower WMC may struggle to keep pace with fast-paced discussions, as they must retrieve and synthesize prior knowledge while maintaining attention to ongoing dialog (Alloway & Alloway, 2010). Structured discussion formats that reduce WM load, such as providing pre-discussion prompts or guided questioning, can enhance meaningful participation for all learners (Kyllonen & Christal, 1990).

3.2.2. Note-Taking and Following Instructions

Note-taking is a cognitively demanding activity that relies heavily on WM. Effective note-taking requires students to process spoken information while summarizing key points in real time. Research indicates that as cognitive load increases, note-taking accuracy can decline, with students capturing fewer key details and struggling to organize their notes effectively. For instance, Peverly et al. (2013) found that handwriting speed, language comprehension, and WM were significantly related to the quality of notes taken during lectures. Similarly, following multi-step instructions requires learners to maintain the previous steps in WM while executing new ones, a process that individuals with WM challenges often find difficult. Strategies such as breaking tasks into simple steps and using clear, specific language can help support students with WM difficulties in following instructions.

3.2.3. Classroom Management and Behavior Regulation

Self-regulation and classroom behavior are also likely influenced by WMC, as students with lower WMC tend to experience more difficulty sustaining attention, resisting distractions, and switching between tasks efficiently (Engle, 2018). These challenges are often linked to inattention and impulsivity, which can affect learners' ability to manage classroom responsibilities and remain engaged with long-duration tasks. Classroom strategies that provide external cognitive supports, such as visual reminders, structured task transitions, and explicit sequencing of activities, can help mitigate these difficulties and enhance self-regulation in lower-WMC students.

Taken together, these findings underscore the fundamental role of WMC in both academic performance and classroom engagement. The evidence consistently highlights that students with higher WMC are better equipped to handle the cognitive demands of academic tasks and classroom activities, while those with lower WMC often face significant challenges. These findings suggest that instructional strategies aimed at reducing cognitive load and providing external support can help mitigate the difficulties faced by lower-WMC students, enhancing both their academic performance and classroom engagement. As

research continues to explore the complex relationship between WMC and learning, it remains clear that fostering WMC through targeted interventions could have a profound impact on educational outcomes.

4. The ATI Framework in Educational Research

The previous sections established the role of WMC in academic performance. This section explores the ATI framework, a conceptual approach designed to examine how individual differences, particularly in cognitive aptitudes such as WMC, influence the effectiveness of instructional strategies. The ATI framework challenges the assumption that a single instructional strategy is universally effective. Instead, it emphasizes that the success of an instructional strategy depends on the learner's cognitive characteristics, suggesting that tailoring instruction to individual aptitudes can optimize learning outcomes (Cronbach & Snow, 1977; Rohrer & Pashler, 2010).

4.1. The Evolution and Importance of the ATI Framework

Early ATI research focused primarily on broad intelligence measures, but recent work has shifted toward examining specific cognitive traits, such as WMC, executive function, and attentional control. These abilities provide a more precise explanation for why certain students benefit from particular instructional strategies while others struggle (Mayer, 2017). A growing body of research supports the role of WMC as a key aptitude in ATI studies. Unlike general intelligence, which encompasses multiple cognitive domains, WMC is a specific, measurable construct that directly influences how students encode, store, and retrieve information in real time. Empirical studies have confirmed that WMC moderates the effectiveness of instructional strategies such as multimedia learning (Sanchez & Wiley, 2009), retrieval practice (Tse & Pu, 2012), and spaced repetition (Bui et al., 2013). These studies demonstrate that instructional strategies optimized for lower-WMC learners, such as reducing extraneous cognitive load or providing structured guidance, do not necessarily hinder higher-WMC learners. This finding has important implications for designing equitable instructional strategies.

However, not all ATI effects are benign or additive. While many studies report ordinal interactions—in which an instructional strategy benefits one group more than another but does not harm either—a smaller but important set of studies demonstrates crossover interactions. In these cases, an instructional method that supports one group may actively impede learning in another. For instance, high levels of instructional guidance or scaffolding may benefit lower-WMC students but reduce performance among higher-WMC learners by limiting opportunities for productive struggle or deeper engagement (Kalyuga, 2007; Sweller, 2017). These crossover patterns are pedagogically significant, as they highlight situations where well-intentioned supports may unintentionally harm subsets of learners. Recognizing such effects is crucial for developing instructional approaches that are not only inclusive but also avoid inadvertently disadvantaging higher-capacity students.

As education continues to evolve toward technology-enhanced and adaptive learning environments, ATI is becoming increasingly relevant. Digital learning platforms now have the capacity to dynamically adjust instructional content based on students' cognitive abilities, effectively implementing ATI principles in real-world settings (Schmidt-Weigand & Scheiter, 2011). By explicitly considering ATI effects, researchers and educators can develop more evidence-based, personalized learning strategies that accommodate a broad spectrum of cognitive strengths and weaknesses.

4.2. Methodological Approaches in ATI Research

To investigate how WMC interacts with instructional strategies, researchers use a variety of methodological approaches. These methods allow for the statistical detection of interactions between aptitude (WMC) and treatment (instructional intervention), helping to determine whether certain instructional strategies benefit some learners more than others.

One of the most widely used techniques in ATI research is moderation analysis, which tests whether the relationship between an instructional strategy and learning outcome changes depending on WMC. For example, Sanchez and Wiley (2006) examined how extraneous visual information in multimedia learning affected high- and low-WMC learners. Using moderation analysis, they found that lower-WMC learners were significantly more distracted by irrelevant images, whereas higher-WMC learners were unaffected. This suggests that removing unnecessary visuals in educational materials benefits lower-WMC learners without harming others, a clear demonstration of an ATI effect. Similarly, Agarwal et al. (2017) applied moderation analyses to study retrieval practice, finding that lower-WMC learners benefited disproportionately from testing when feedback was provided, whereas higher-WMC learners showed more stable performance across conditions. These findings highlight the importance of considering WMC when designing instructional strategies that rely on retrieval practice.

While moderation analyses are effective for detecting interaction effects, they often require researchers to dichotomize continuous traits like WMC into "high" or "low" categories. This not only reduces statistical power but also oversimplifies the nature of individual differences. A more advanced approach involves structural equation modeling (SEM) and latent variable interaction models, which preserve the continuous nature of cognitive traits and account for measurement error in both aptitude and outcome variables. By modeling WMC as a latent construct based on multiple observed indicators, SEM provides more precise estimates of aptitude–treatment interactions than traditional moderation techniques.

Latent variable interaction models extend this precision further by capturing the subtle ways in which individual differences in WMC influence responsiveness to instruction. These models allow researchers to test whether the effect of an instructional strategy varies across the full range of WMC, rather than relying on arbitrary groupings. For instance, recent work has shown that learners with high attentional control—a construct often correlated with WMC—may not benefit from the same supports (e.g., scaffolding or retrieval cues) that assist learners with lower capacity (Unsworth et al., 2021). Although statistically complex, SEM and latent variable approaches offer richer, more accurate representations of cognitive aptitude, making them essential tools for developing personalized, empirically grounded instructional strategies.

4.3. Practical Considerations in ATI Research

While ATI research provides valuable insights into cognitive learning processes, several methodological and practical considerations must be addressed. One key challenge is ensuring the accurate measurement of WMC, as not all WM tasks capture the same cognitive processes. Many ATI studies rely on complex span tasks, such as operation span, reading span, and symmetry span, to assess WMC. While these tasks are reliable measures of WMC, some, like operation span, are more domain-general and may not be as effective at predicting specific academic outcomes, such as reading comprehension. For example, the reading span task, which involves holding and integrating verbal information, is better suited for predicting reading comprehension performance. Furthermore, some simple span tasks, like digit span, do not measure attentional control, a crucial component of WMC (Redick et al., 2016). Researchers must carefully select validated WM assessments that align

with the specific cognitive processes relevant to their study in order to ensure the precise measurement of aptitude differences.

Another challenge in ATI research is accounting for additional factors that influence instructional effectiveness, such as prior knowledge, motivation, and self-regulation skills. Failing to control for these variables can lead to spurious ATI effects. For example, a student with low WMC but high motivation may compensate for WM deficits by employing greater effort and engagement, potentially masking the true interactions between WMC and instructional strategies (Sweller, 2017). Similarly, a student with low WMC but strong self-regulation skills might manage cognitive demands more effectively than a student with similarly low WMC but poor self-regulation, thus obscuring the real relationship between WMC and instructional outcomes (Duckworth & Seligman, 2005). To provide a clearer understanding of how WMC interacts with instructional strategies, future research should carefully control for these confounding variables to more accurately assess their impact on learning.

Many ATI studies are conducted in controlled laboratory environments rather than authentic educational settings. While these experiments provide valuable theoretical insights, their real-world applicability remains limited. Future research should investigate how ATI effects translate to classroom settings, online learning environments, and collaborative learning scenarios to enhance practical relevance. Additionally, most ATI studies treat WMC as a stable, trait-level characteristic, but emerging research suggests that WMC can fluctuate based on factors like fatigue, stress, and cognitive load (Ilkowska & Engle, 2010). Exploring whether temporary changes in WMC influence ATI effects would offer a more dynamic understanding of how instructional treatments interact with cognitive capacity.

In summary, the ATI framework offers a compelling model for understanding how instructional strategies interact with individual differences in WMC. By employing rigorous methodological techniques, researchers can gain deeper insights into how ATI effects shape learning outcomes. However, to maximize the real-world applicability of these findings, future studies should consider state-dependent changes in WMC, explore instructional interactions in authentic educational contexts, and refine methods for adaptive instructional design. Understanding how WMC moderates instructional effectiveness is crucial for developing more inclusive, evidence-based educational practices that accommodate diverse learners.

5. WM and Instructional Strategies: An ATI Perspective

Understanding how instructional strategies interact with WMC is central to optimizing learning outcomes for diverse student populations. While effective instructional strategies have been identified through cognitive and educational research, their impact often depends on the learner's cognitive abilities. The ATI framework provides a lens through which to examine how different instructional strategies benefit students based on their WMC. This section discusses key instructional approaches, reviewing how they interact with WMC and providing evidence-based recommendations for educational practice.

5.1. Multimedia Learning

Multimedia instruction—integrating verbal, visual, and auditory materials—places considerable demands on WM, as learners must coordinate multiple streams of information. According to the cognitive load theory (CLT; Sweller, 1988), multimedia materials can either facilitate or hinder learning, depending on how they interact with WMC. Studies indicate that lower-WMC learners are particularly vulnerable to extraneous cognitive load in multimedia learning environments. Sanchez and Wiley (2006) found that unnecessary visuals impaired comprehension for lower-WMC learners but had minimal effect on higher-

WMC learners. Similarly, Lusk et al. (2009) demonstrated that segmenting multimedia presentations benefited lower-WMC individuals by reducing information overload. To optimize multimedia learning, instructional designers should limit extraneous elements, such as decorative images or animations, and provide segmentation, allowing students to control the pace of content delivery. Additionally, dual-channel processing—where spoken words are paired with relevant visuals instead of redundant on-screen text—has been shown to reduce cognitive demands for lower-WMC learners while maintaining benefits for higher-WMC learners (Mayer & Moreno, 2003).

While this review centers on ATIs, many of the patterns described here are consistent with CLT. For example, interventions that benefit lower-WMC learners—such as reducing visual clutter or enabling self-paced navigation—function by reducing extraneous cognitive load. This alignment suggests that CLT offers a complementary perspective on why certain instructional strategies interact with learner aptitudes, particularly in contexts where managing multiple information sources taxes limited cognitive resources.

5.2. Test-Enhanced Learning (Retrieval Practice)

The retrieval practice effect—the process of recalling learned information to strengthen memory—has been widely recognized as one of the most effective learning strategies (Roediger & Butler, 2011). However, research suggests that WMC moderates the benefits of retrieval practice, particularly in high-stakes or anxiety-inducing learning contexts. Tse and Pu (2012) found that while retrieval practice benefited all learners, the effectiveness of repeated testing was less pronounced for individuals with lower WMC and higher test anxiety. Specifically, lower-WMC learners with high test anxiety showed more errors and less benefit from testing compared to higher-WMC learners. Brewer and Unsworth (2012) found that retrieval practice benefits learners regardless of WMC, but that lower-ability learners may experience greater overall gains due to their poorer baseline memory performance. Agarwal et al. (2017) further demonstrated that retrieval practice is especially beneficial for lower-WMC learners when cognitive load is minimized. These findings suggest that retrieval practice should be structured to reduce cognitive demands for lower-WMC learners, particularly in contexts that may induce test anxiety, by incorporating low-stakes quizzes, cumulative assessments, and immediate feedback.

5.3. Spaced Learning (Distributed Practice)

The spacing effect—the phenomenon where information retention improves when learning is distributed over time rather than massed into a single session—has been extensively studied in cognitive psychology (Cepeda et al., 2006). Research suggests that the optimal spacing interval may depend on individual differences such as working memory capacity (WMC). For instance, Bui et al. (2013) found that individuals with higher WMC benefited more from difficult intervening tasks between repetitions, while those with lower WMC showed better retention with easier intervening tasks. This indicates that lower-WMC learners may struggle with complex tasks during longer intervals due to difficulties in managing cognitive load. Similarly, Delaney et al. (2017) examined whether WMC influences the spacing effect in cued recall and found that while both WMC and spacing independently improved memory performance, they did not significantly interact—suggesting that spacing benefits all learners regardless of WMC. To optimize spaced learning, educators should consider adjusting the difficulty of intervening tasks based on individual WMC levels. Adaptive learning platforms can incorporate dynamic scheduling that tailors task complexity and spacing intervals to match learners' cognitive capacities, thereby enhancing learning outcomes.

5.4. Interleaved Practice

The interleaving effect—alternating between different types of problems or concepts—has been shown to enhance long-term retention and problem-solving skills compared to blocked practice (Rohrer & Taylor, 2007). However, its effectiveness varies based on WMC. Sana et al. (2018) directly investigated whether WMC moderates the interleaving effect and found that interleaved practice benefits learners regardless of their WMC, suggesting that its advantages are broadly applicable. Similarly, Guzman-Munoz (2016) examined how WMC influences the effectiveness of interleaved inductive learning and found that while interleaving benefits all learners, those with higher WMC may find it easier to process the mixed examples. Suzuki et al. (2020) also explored WMC's role in proceduralizing L2 syntax through interleaved grammar practice, suggesting that cognitive resources influence the effectiveness of interleaving in certain contexts. Additionally, Sana et al. (2017) demonstrated that study sequence impacts inductive learning of cognitive concepts, further supporting the instructional value of interleaved practice. Given these findings, educators should integrate interleaved practice into instructional design, particularly in ways that promote conceptual differentiation and retrieval-based learning. Structured worksheets, mixed-problem assignments, and varied practice sessions can help ensure that all students benefit from this evidence-based strategy.

5.5. Worked Examples and Explicit Instruction

Worked examples—step-by-step demonstrations of problem-solving procedures—are particularly beneficial for learners with lower working memory capacity (WMC), as they reduce cognitive load by minimizing the need for simultaneous problem-solving and information retention (Sweller et al., 2019). According to CLT, worked examples effectively manage cognitive demands, especially for novice learners. Research indicates that lower-WMC learners benefit more from worked examples compared to unstructured problem-solving approaches, as worked examples help scaffold complex information and prevent cognitive overload (Sweller, 2017; Paas & van Merriënboer, 1994). The expertise reversal effect further suggests that as learners develop proficiency, the effectiveness of worked examples diminishes, allowing higher-WMC learners to transition to independent problem-solving earlier without negative effects (Kalyuga, 2007). Similarly, Renkl (2014a, 2014b) highlights that gradually fading worked examples—transitioning from fully guided solutions to problem-solving tasks—is an effective instructional strategy that accommodates learners as they develop expertise. For instructional application, educators should consider this gradual fading process to ensure that lower-WMC learners are not overwhelmed while still promoting active engagement.

5.6. Instructional Guidance: Inquiry vs. Direct Instruction

The debate between inquiry-based learning and direct instruction is particularly relevant in the context of working memory (WM). While higher-WMC learners can often manage the open-ended nature of inquiry-based learning, lower-WMC learners may struggle with the cognitive demands of self-directed discovery. Kirschner et al. (2006) found that lower-WMC students benefited significantly more from explicit instruction compared to inquiry-based approaches, as direct guidance reduces extraneous cognitive load. Similarly, Ginns (2005) highlighted that structured guidance, such as worked examples, helps learners better retain information by managing cognitive load effectively. However, well-scaffolded inquiry learning, where guidance is gradually removed as learners develop expertise, can mitigate WM-related challenges. To optimize learning, educators should provide structured guidance early on and then gradually transition to inquiry-based methods as students build expertise.

5.7. Flipped Classrooms

Flipped classrooms—where students engage with instructional materials before class and apply concepts during in-class activities—can influence cognitive load and accommodate different WMC levels. Diningrat et al. (2023) investigated the effect of an extended flipped classroom model on students' reading comprehension and found that while flipped learning was beneficial, students with high WMC outperformed those with lower WMC. This suggests that flipped classrooms can be effective but must be carefully structured to support lower-WMC learners. Bergmann and Sams (2012) emphasized that flipped classrooms reduce cognitive load by allowing students to learn at their own pace, which may be particularly helpful for learners who struggle with memory retention. However, to maximize effectiveness, flipped learning must ensure that students complete pre-class work and receive structured guidance during in-class activities.

5.8. Feedback (Types and Delivery)

Feedback plays a crucial role in learning, but its effectiveness depends on WMC. Li (2013) investigated the interactions between implicit and explicit feedback and individual differences in language analytic ability and WM in second language learning. The study found that learners with higher WMC benefited more from explicit feedback, while those with lower WMC showed comparable gains from both implicit and explicit feedback. Similarly, Fyfe et al. (2015) examined the role of WMC in the effectiveness of different feedback types in mathematical problem-solving. Their research indicated that lower-WMC learners benefited more from outcome feedback, which provides information on the correctness of an answer, whereas higher-WMC learners benefited similarly from both outcome and strategy feedback, the latter offering guidance on problem-solving strategies. These findings are consistent with the broader feedback literature, which suggests that different types of feedback can optimize learning based on cognitive capacity (Hattie & Timperley, 2007). To optimize learning, feedback should be tailored based on cognitive load considerations, providing structured hints and scaffolding for lower-WMC learners while allowing more open-ended correction for higher-WMC students.

In summary, this section has highlighted how WMC may influence the effectiveness of different instructional strategies. While direct empirical studies on WMC–instructional strategy interactions remain limited, findings from CLT and related research suggest that lower-WMC learners often benefit from structured, scaffolded, and guided approaches that help manage cognitive demands. In contrast, higher-WMC learners may be better equipped to engage with more autonomous and generative learning environments that require greater self-regulation and cognitive flexibility. However, given the limited number of studies directly examining these interactions, further research is needed to refine our understanding of how WMC shapes learning outcomes across different instructional contexts. The following section will explore the interplay between metacognition, retrieval cues, and WMC, considering how learning strategies can be adapted to accommodate individual cognitive differences and optimize retention and transfer.

6. WM, Metacognition, and Retrieval Cues

While instructional treatments can be designed to accommodate differences in WMC, another critical factor influencing learning is metacognition—the ability to monitor and regulate one's own cognitive processes. Metacognitive skills enable learners to assess their understanding, adjust their study strategies, and retrieve relevant information efficiently. However, the effectiveness of metacognitive strategies is closely tied to WMC, as individuals with lower WMC often struggle with self-monitoring and controlled retrieval processes. This section explores the relationship between WMC, metacognition, and re-

trieval cues, offering insights into how instructional design can better support learners with varying WMC.

6.1. Metacognitive Monitoring and Strategies

Metacognitive monitoring—the ability to evaluate one's level of understanding and adjust learning strategies accordingly—is essential for academic success. However, research suggests that lower-WMC learners tend to be less accurate in monitoring their comprehension and predicting their performance compared to higher-WMC learners (Griffin et al., 2008). This discrepancy arises because effective metacognitive monitoring requires holding multiple pieces of information in WM while simultaneously evaluating their coherence and accuracy, a process that is particularly demanding for those with limited WM resources. Griffin et al. (2008) found that lower-WMC learners often overestimate their comprehension and fail to recognize gaps in their understanding. However, this monitoring deficit can be mitigated through structured metacognitive prompts. When participants were prompted to self-explain key concepts after reading, their comprehension monitoring accuracy improved, effectively reducing WMC-related performance gaps. Similarly, rereading the text before making comprehension judgments helped lower-WMC learners improve their monitoring accuracy, suggesting that reducing concurrent WM demands allows for more accurate self-assessment. These findings are consistent with the broader metacognitive literature, which suggests that students with higher WMC tend to have better metacognitive awareness and strategy adjustment abilities (Koriat, 2007; Veenman et al., 2006).

From an instructional standpoint, embedding metacognitive scaffolds can enhance self-monitoring skills for lower-WMC learners. Strategies such as prompting learners to generate explanations, summarize key points, or predict test performance can help them regulate their learning more effectively. Educators can also train students in metacognitive strategies explicitly, teaching them to pause periodically, assess their understanding, and adjust their study methods accordingly. It is possible that self-regulated learning (SRL) training disproportionately benefits lower-WMC learners. By incorporating goal-setting, self-questioning, and monitoring strategies, lower-WMC individuals can develop compensatory mechanisms that allow them to allocate cognitive resources more effectively. This suggests that explicit instruction in metacognitive regulation should be integrated into curricula, particularly for students who struggle with attentional control and self-monitoring. These findings are echoed by studies on cognitive load and self-regulation, suggesting that explicitly teaching metacognitive strategies can mitigate the cognitive challenges faced by lower-WMC learners (e.g., Dunlosky & Metcalfe, 2009; Schraw, 2006).

6.2. Retrieval Cues and Search Processes in WM

Retrieval cue selection is a critical mechanism that influences learning efficiency. WM plays a pivotal role in guiding controlled searches of Long-Term Memory (LTM), but lower-WMC learners often struggle with selecting appropriate retrieval cues, leading to increased interference from irrelevant information (Unsworth & Engle, 2007). Unsworth et al. (2012a) found that lower-WMC individuals were more likely to experience retrieval failure due to ineffective cue selection. However, providing structured retrieval cues, such as category labels, contextual hints, or partial word cues, significantly improved recall performance for these learners. This suggests that educators should incorporate retrieval scaffolding into learning materials to help lower-WMC students retrieve information more efficiently.

Classroom assessments can be designed to support retrieval for lower-WMC learners by incorporating scaffolded recall questions with keywords or conceptual categories to reduce retrieval failures. Concept maps and structured outlines can also serve as external retrieval cues, helping students organize and recall information. Additionally, cumulative

assessments that revisit prior concepts can reinforce memory recall pathways and make retrieval a habitual process. In fact, lower-WMC learners seem to benefit from frequent low-stakes quizzes with guided retrieval cues, which strengthen recall pathways and reduce reliance on inefficient search strategies (Agarwal et al., 2017; Unsworth & Engle, 2007). This aligns with research on spaced retrieval practice, suggesting that gradually introduced retrieval cues can help build strong memory networks for lower-WMC learners.

Tailoring instructional strategies to support metacognitive and retrieval processes is crucial for optimizing learning outcomes. By incorporating techniques like metacognitive training, structured retrieval cues, and spaced retrieval practice, educators can better support lower-WMC learners and bridge performance gaps. Continued research into adaptive strategies will further enhance personalized learning, helping students of all cognitive abilities reach their full potential.

7. Integrating ATI Research into Educational Practice

The ATI framework provides valuable insights into how instructional strategies can be tailored to accommodate individual differences in WMC. While ATI research has traditionally been conducted in controlled laboratory settings, the challenge now lies in translating these findings into real-world educational applications. This section explores practical approaches for implementing ATI-based strategies in classroom instruction, assessment, and adaptive learning environments while addressing key challenges and research considerations for future investigations.

7.1. Guidelines for Educators

One of the most effective ways to integrate ATI principles into education is through adaptive instruction, in which instructional content dynamically adjusts based on learners' cognitive characteristics. Advances in digital learning technologies and intelligent tutoring systems allow for the real-time assessment of individual cognitive abilities, enabling instructional materials to be personalized accordingly. For example, adaptive learning platforms can measure students' WMC using embedded cognitive tasks and then modify instructional complexity based on their processing capacity. Research suggests that such adaptive learning environments improve engagement and retention by ensuring that learners are neither overwhelmed nor under-stimulated (Kalyuga & Plass, 2017; Paas & van Merriënboer, 2020). Higher-WMC learners can be presented with more complex, self-guided challenges, while lower-WMC learners receive structured, scaffolded instruction to manage cognitive load effectively (Kalyuga, 2009).

Differentiated instruction provides another avenue for applying ATI principles in classrooms. Unlike adaptive learning, which relies on automated adjustments, differentiated instruction requires teachers to proactively tailor lesson plans, assessments, and instructional strategies to meet diverse cognitive needs. Educators can use informal assessments of WMC—such as students' ability to follow multi-step instructions, maintain attention, and recall previously learned material—to determine which students require additional cognitive support. For lower-WMC learners, instructional strategies should emphasize breaking down complex tasks into smaller, more manageable steps, incorporating guided practice with worked examples, and providing external memory aids (Swanson et al., 2013). In contrast, higher-WMC learners tend to benefit from instructional techniques that encourage independent problem-solving, such as inquiry-based learning, interleaved practice, and self-explanation (Van Merriënboer & Sweller, 2010).

The ATI framework also aligns with universal design for learning (UDL), an educational approach that promotes accessibility for all students by offering multiple means of engagement, representation, and expression. By integrating principles of the cognitive

load theory, UDL allows students with varying WMC levels to learn through different modalities without being segregated into separate instructional tracks (Meyer et al., 2014). For example, multimedia instruction can incorporate both text-based and auditory explanations, enabling students to select the modality that best suits their cognitive processing capacity (Mayer, 2017). Similarly, multiple formats for demonstrating knowledge—such as written responses, verbal explanations, or concept mapping—can help lower-WMC learners express their understanding more effectively.

A crucial pedagogical strategy within ATI-informed instruction is scaffolding, which provides structured support at the beginning of the learning process and gradually reduces assistance as students gain proficiency. Scaffolding techniques, including guided questioning, explicit modeling, and structured problem breakdowns, help lower-WMC learners manage cognitive demands more effectively (Van de Pol et al., 2010). Research suggests that lower-WMC learners struggle with open-ended problem-solving tasks due to increased cognitive load, whereas structured guidance helps them acquire and apply new information more efficiently (Kirschner et al., 2006). In contrast, higher-WMC learners often benefit from reduced instructional guidance, as they can engage in deeper cognitive processing and self-directed learning (Clark et al., 2012). Therefore, educators should gradually transition students from explicit instruction to independent problem-solving, ensuring that instructional demands align with individual cognitive capacities.

7.1.1. Assessment and Classroom Collaboration in ATI-Based Instruction

Assessment design plays a critical role in ensuring that ATI principles are applied effectively. Traditional assessments often assume that all students process and retrieve information similarly, yet research suggests that WMC influences test performance independently of actual content knowledge (Unsworth & Robison, 2020). To mitigate this, assessments should be structured to minimize unnecessary WM demands for lower-WMC learners. For example, cued recall questions, which provide prompts or structured retrieval cues, can help students access relevant information without overwhelming their cognitive resources (Unsworth et al., 2012a). Similarly, extended testing time or structured cognitive breaks can enhance performance by reducing stress-induced WM depletion (Alloway & Alloway, 2010). In contrast, higher-WMC learners can be challenged with complex, integrative assessments that require synthesizing information across multiple domains (Tarchi et al., 2024).

Classroom collaboration and peer-assisted learning also offer opportunities for ATI-informed instruction. Research has shown that structured peer collaboration benefits lower-WMC learners by distributing cognitive demands among multiple individuals, effectively reducing the cognitive load on any single learner (Kirschner et al., 2018). However, collaborative learning environments must be carefully structured to ensure that lower-WMC learners actively contribute rather than passively relying on higher-WMC peers. Techniques such as reciprocal teaching, where students take turns explaining concepts, and jigsaw learning, where each student is responsible for mastering a different subtopic before teaching it to peers, can promote engagement while balancing cognitive demands (Palincsar & Brown, 1984; Rosenshine & Meister, 1994).

7.1.2. Adapting Group Instruction Without Stigmatization

While adaptive and differentiated instruction are well suited to individualized digital learning environments, many educational settings rely on group-based instruction. Applying ATI principles in a classroom context without stigmatizing students with lower WMC requires a multi-faceted approach. One effective method is to incorporate universal design for learning (UDL), which provides multiple means of engagement, representation,

and expression. For instance, a lesson can simultaneously include verbal explanations, visual diagrams, and interactive activities, allowing students to select the modality that best aligns with their cognitive processing capacity. Such an approach ensures that learners benefit from differentiated supports without being visibly labeled.

In addition to UDL, flexible grouping strategies can help mitigate stigmatization. Rather than categorizing students by ability, educators can adopt dynamic grouping, where students are periodically reorganized based on task requirements or instructional goals. This approach promotes collaborative learning without permanently isolating lower-WMC learners. For example, in problem-solving tasks, groups can be intentionally composed of diverse learners, enabling peer support without overtly highlighting individual differences.

Layered scaffolding offers another practical solution. Rather than explicitly categorizing students, educators can make optional supports universally available, such as guided practice, visual cues, or step-by-step instructions. Lower-WMC learners can choose to access these supports without being singled out, while higher-WMC learners can engage in independent problem-solving. This self-selected support model allows all students to engage at an appropriate cognitive level without stigma.

Finally, instructor sensitivity and professional development are critical. Educators should be trained to recognize diverse cognitive needs without relying on explicit labels or making assumptions. By adopting a strength-based perspective—emphasizing that some learners benefit from structured practice while others thrive with exploratory tasks—teachers can maintain an inclusive, respectful classroom culture. Collectively, these strategies ensure that ATI principles can be effectively implemented in group instruction without risking negative social perceptions, allowing all learners to benefit from cognitively aligned support.

7.2. Guidelines for Researchers

While Section 7.1 focused on how educators can apply ATI principles in instructional settings, this section offers complementary guidance for researchers. We highlight methodological and design considerations for studying aptitude–treatment interactions and suggest empirical strategies to improve the rigor and ecological validity of future research.

From a research perspective, the ATI framework presents an opportunity to refine methodologies for examining how cognitive factors influence learning. However, many existing studies lack the statistical rigor required to fully capture these interactions. Moderation analyses remain the most commonly used method for testing ATI effects, allowing researchers to assess how WMC influences the effectiveness of various instructional strategies. Structural equation modeling (SEM) and latent variable interaction models offer additional statistical tools that can provide a more precise understanding of how cognitive traits moderate learning outcomes across different instructional conditions.

Practical considerations must also be taken into account when measuring WMC in ATI research. Traditional WM tasks, such as complex span tasks, remain the gold standard for assessing WMC, but researchers must ensure that these measures are both reliable and ecologically valid in instructional contexts. Many studies rely on laboratory-based assessments, but these tasks may not fully capture how WMC operates in real-world educational settings (Holmes & Gathercole, 2014). Additionally, small sample sizes can obscure ATI effects, as these interactions are often subtle. To enhance statistical power, studies should be designed with sufficient sample sizes and robust analytic techniques to detect meaningful interactions.

Moreover, longitudinal research is needed to determine the sustained impact of ATI-based instructional approaches. Most ATI studies examine short-term learning gains, but tracking students' cognitive performance over time would provide more robust evidence of

how WMC interacts with instructional treatments across different stages of learning. Longitudinal designs would also allow researchers to investigate whether repeated exposure to ATI-aligned instruction can help lower-WMC learners develop compensatory strategies, potentially reducing performance gaps over time.

A key challenge in ATI research is ensuring ecological validity, as many studies are conducted in controlled laboratory settings rather than real-world educational environments. Future research should prioritize classroom-based interventions, examining how WMC interacts with instructional strategies in authentic learning environments. Additionally, expanding ATI research beyond traditional academic domains—such as mathematics and reading comprehension—to emerging instructional paradigms like flipped classrooms, virtual learning environments, and AI-driven tutoring systems will provide a more comprehensive understanding of how ATI principles apply in diverse educational settings.

Another major limitation of current ATI research is the lack of direct cognitive assessments in classroom settings. While standardized WM tasks such as complex span tasks are used in research, they are not commonly integrated into everyday educational practice. Instead, educators must rely on indirect indicators of WMC, such as attention regulation, task persistence, and instructional responsiveness (Gathercole & Alloway, 2008). Developing efficient, classroom-friendly cognitive assessment tools would allow for more precise tailoring of instructional strategies based on individual cognitive capacities.

The ATI framework offers a powerful model for designing instructional strategies that align with students' cognitive abilities, particularly WMC. By incorporating adaptive learning, differentiated instruction, scaffolding, and structured assessment design, educators can create more equitable and effective learning environments. While challenges remain in scaling ATI-based instruction, advances in educational technology and adaptive learning models provide promising pathways for widespread implementation. Future research should continue refining statistical methodologies, expanding ATI investigations across different instructional contexts, and exploring interventions aimed at enhancing WMC. As the field progresses, bridging the gap between cognitive science and educational practice will be essential in ensuring that ATI-based strategies translate into meaningful improvements in student learning outcomes.

8. Future Directions for ATI and WM in Education

In addition to methodological recommendations, there are several conceptual and applied frontiers that merit further investigation. This final section outlines unresolved questions and proposes future research directions spanning instructional strategy design, cognitive states, learner diversity, and emerging educational technologies.

Despite significant progress in understanding how WMC interacts with instructional strategies, many questions remain unresolved. The ATI framework has provided valuable insights, but further research is needed to refine our understanding of how cognitive traits shape learning outcomes. This section highlights key areas for future investigation, including under-explored instructional strategies, contextual factors influencing WMC, and the implications of ATI research for special populations.

One of the most pressing gaps in ATI research is the limited investigation of analogy learning and its interaction with WMC. Analogies are powerful instructional tools that facilitate knowledge transfer by mapping familiar concepts onto new material, aiding conceptual understanding in subjects like science, mathematics, and second language acquisition (Richland et al., 2007). However, it remains unclear whether the cognitive demands of analogy learning differentially impact students based on their WMC. On one hand, analogies may reduce cognitive load for lower-WMC learners by providing structured knowledge frameworks, but on the other hand, constructing and applying analogies requires abstract

reasoning and relational integration, which may disproportionately burden learners with lower WMC (Richland & Burchinal, 2013). Future research should explore how different types of analogies—simple, complex, self-generated, or instructor-provided—affect learners across the WMC spectrum.

Similarly, spaced learning interventions require further refinement within an ATI framework. While spacing has been consistently shown to enhance retention (Cepeda et al., 2006; Roediger & Butler, 2011), the optimal spacing interval for learners with varying WMC remains an open question. Some evidence suggests that higher-WMC learners benefit from longer intervals between study sessions due to their superior ability to retrieve and integrate previously learned information, whereas lower-WMC learners may require shorter, more frequent exposures to prevent information loss (Bui et al., 2013). However, current research is largely correlational, and few studies have directly manipulated spacing conditions based on individual WMC. Future research should systematically examine whether adjusting spacing intervals based on WMC leads to more effective retention and knowledge transfer.

Another promising but controversial area of research is WM training and its potential impact on learning outcomes. While early studies suggested that WM training could lead to broad cognitive improvements (Jaeggi et al., 2008), more recent meta-analyses indicate that such gains are typically limited to the trained task, with minimal transfer to academic performance or real-world problem-solving (Melby-Lervåg et al., 2016). However, it remains unclear whether specific populations—such as students with WM deficits or learning disabilities—might derive greater benefits from targeted training interventions. Future research should explore whether combining WM training with ATI-informed instructional techniques could produce more meaningful improvements in learning outcomes, particularly for at-risk student populations.

Beyond instructional techniques, contextual and situational factors play a crucial role in moderating WMC's influence on learning. While WMC is generally considered a stable trait, evidence suggests it can fluctuate based on stress, fatigue, anxiety, and environmental distractions (Ilkowska & Engle, 2010). This raises important questions about whether ATI-based instructional strategies should be adapted based on learners' transient cognitive states. For example, learners with high WMC may perform well under optimal conditions but experience cognitive overload when faced with high-stakes testing or information-dense lectures (Vogel & Schwabe, 2016). Conversely, lower-WMC learners may benefit from interventions that minimize situational stressors, such as structured retrieval practice or externally provided organizational cues (Putnam et al., 2016). More research is needed to determine how state-dependent variations in WMC interact with instructional treatments and how educators can accommodate these fluctuations.

Developmental factors must also be considered. WMC changes significantly over the lifespan, with younger learners typically showing lower and more variable capacity. These developmental differences suggest that ATI effects may be particularly pronounced in childhood, and instructional strategies may need to be adapted accordingly. For example, younger learners may benefit from more sustained scaffolding, whereas older learners may benefit from the gradual removal of supports.

Another particularly important but underdeveloped area of ATI research concerns special populations, including students with learning disabilities, ADHD, and neurological disorders. Many ATI studies focus on general populations, yet learners with specific cognitive impairments may exhibit fundamentally different interactions with instructional strategies (Gathercole & Alloway, 2008). For example, students with ADHD often exhibit reduced WMC and impaired attentional control, which may alter the effectiveness of interventions designed to support lower-WMC learners (Martinussen & Major, 2011).

Similarly, students with dyslexia may struggle with instructional methods that rely heavily on phonological WM, whereas students with autism spectrum disorder (ASD) may show differential responses to structured versus open-ended instructional designs (Pellicano, 2012). Tailoring ATI-based strategies to meet the needs of diverse learners is a critical next step for research in this area.

Another important area for future research is the role of motivation and self-regulated learning strategies within ATI-based instruction. Students with lower WMC often face additional challenges related to task persistence, effort allocation, and self-monitoring, which can influence how well they respond to instructional strategies designed to accommodate cognitive load differences (Dunlosky & Thiede, 2013). Metacognitive training—including techniques such as self-explanation, reflection prompts, and guided goal-setting—may serve as a buffer against cognitive overload by helping lower-WMC learners develop compensatory strategies (Bjork et al., 2013). Future studies should investigate how motivational factors interact with cognitive aptitudes in shaping learning outcomes, particularly in self-paced learning environments where students must regulate their own study habits.

Future research should also consider the role of technology and adaptive learning environments in delivering ATI-informed instruction. With the rise in artificial intelligence-driven educational platforms, there is an opportunity to develop adaptive systems that dynamically adjust instructional complexity based on real-time assessments of learners' cognitive capacity (Koedinger et al., 2012). Such systems could personalize learning experiences by modulating cognitive load, providing targeted scaffolding, or recommending optimal retrieval practice intervals (Kalyuga, 2007). However, rigorous empirical validation is necessary to determine the effectiveness of these approaches across different learner populations.

9. Conclusions

This paper has explored the critical role of WMC in shaping how learners respond to instructional treatments, emphasizing the importance of integrating the ATI framework into educational research and practice. Decades of research in cognitive psychology and education have established that individual differences in WMC significantly influence learning outcomes across a wide range of academic domains, including reading comprehension, mathematics, second language acquisition, and science reasoning (Sweller, 2017; Unsworth et al., 2021). These findings underscore the necessity of moving beyond one-size-fits-all instructional approaches and instead tailoring teaching methods to align with learners' cognitive abilities.

Beyond its implications for educators, this research also highlights important considerations for researchers designing future ATI studies. Future research should focus on how ATI principles can be leveraged in modern, technology-enhanced learning environments, including digital classrooms, virtual reality-based training, and AI-powered adaptive tutoring systems. Investigating how ATI principles can be incorporated into emerging technologies will be crucial for advancing the practical applicability of this research.

Ultimately, this paper underscores the need for continued research and the practical application of ATI principles in education. By systematically considering how WMC influences learning, both researchers and educators can contribute to a more nuanced and effective understanding of how instructional strategies can be tailored to meet the needs of all learners. The goal is not simply to identify which methods work best but to understand for whom and under what conditions they work best, ensuring that education becomes a truly personalized and impactful experience for every student.

Author Contributions: Conceptualization, F.S. and B.F.; data curation, F.S.; writing—original draft preparation, F.S. and B.F.; writing—review and editing, F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Social Sciences and Humanities Research Council, grant number #435-2021-0426 awarded to F.S.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were collected.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ATI	aptitude–treatment interaction
WM	working memory
WMC	working memory capacity

References

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L., III. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory*, *25*(6), 764–771. [CrossRef] [PubMed]

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29. [CrossRef] [PubMed]

Alloway, T. P., & Copello, E. (2013). *Working memory: The connected intelligence*. Psychology Press. [CrossRef]

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. [CrossRef]

Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. International Society for Technology in Education.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*(1), 417–444. [CrossRef] [PubMed]

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language*, *66*(3), 407–415. [CrossRef]

Bui, D. C., Maddox, G. B., & Balota, D. A. (2013). The roles of working memory and intervening task difficulty in determining the benefits of repetition. *Psychonomic Bulletin & Review*, *20*(2), 341–347. [CrossRef]

Burgess, G. C., Gray, J. R., Conway, A. R. A., & Braver, T. S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *Journal of Experimental Psychology: General*, *140*(4), 674–692. [CrossRef]

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. [CrossRef]

Clark, R. E., Kirschner, P. A., & Sweller, J. (2012). Putting students on the path to learning: The case for fully guided instruction. *American Educator*, *36*(1), 5–11. Available online: https://eric.ed.gov/?id=EJ971752 (accessed on 10 May 2025).

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158–1170. [CrossRef]

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. Irvington Publishers.

Delaney, P. F., Godbole, N. R., Holden, L. R., & Chang, Y. (2017). Working memory capacity and the spacing effect in cued recall. *Memory*, *26*(6), 784–797. [CrossRef] [PubMed]

Diningrat, S. W. M., Setyosari, P., Ulfa, S., & Widiati, U. (2023). The effect of an extended flipped classroom model for fully online learning and its interaction with working memory capacity on students' reading comprehension. *Journal of New Approaches in Educational Research*, *12*(1), 78–95. [CrossRef]

Draheim, C., Hicks, K. L., & Engle, R. W. (2016). Combining reaction time and accuracy: The relationship between working memory capacity and task switching as a case example. *Perspectives on Psychological Science*, *11*(1), 133–155. [CrossRef] [PubMed]

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, *16*(12), 939–944. [CrossRef] [PubMed]

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Sage.

Dunlosky, J., & Rawson, K. A. (2019). How cognitive psychology can inform evidence-based education reform. In J. Dunlosky, & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 1–14). Cambridge University Press.

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, *24*, 58–61. [CrossRef]

Engle, R. W. (2018). Working memory and executive attention: A revisit. *Perspectives on Psychological Science*, *13*(2), 190–193. [CrossRef]

Finn, A. S., Kraft, M. A., West, M. R., Leonard, J. A., Bish, C. E., Martin, R. E., Sheridan, M. A., Gabrieli, C. F., & Gabrieli, J. D. (2015). Cognitive skills, student achievement tests, and schools. *Psychological Science*, *26*(6), 736–749. [CrossRef]

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*(2), 226–236. [CrossRef]

Friso-van den Bos, I., van der Ven, S. H., Kroesbergen, E. H., & van Luit, J. E. (2015). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review*, *10*, 29–44. [CrossRef]

Fung, W., & Swanson, H. L. (2017). Working memory components that predict word problem solving: Is it merely a function of reading, calculation, and fluid intelligence? *Memory & Cognition*, *45*(5), 804–823. [CrossRef]

Fyfe, E. R., DeCaro, M. S., & Rittle-Johnson, B. (2015). When feedback is cognitively demanding: The importance of working memory capacity. *Instructional Science*, *43*(1), 73–91. [CrossRef]

Fyfe, E. R., & Rittle-Johnson, B. (2017). Mathematics practice without feedback: A desirable difficulty in a classroom setting. *Instructional Science*, *45*(2), 177–194. [CrossRef]

Gathercole, S. E., & Alloway, T. P. (2008). *Working memory and learning: A practical guide for teachers*. Sage Publications.

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2017). Mathematical cognition deficits in children with learning disabilities and persistent low achievement: A five-year longitudinal study. *Journal of Educational Psychology*, *109*(5), 680–693. [CrossRef] [PubMed]

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction*, *15*(4), 313–331. [CrossRef]

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, *36*(1), 93–103. [CrossRef]

Guzman-Munoz, F. J. (2016). The advantage of mixing examples in inductive learning: A comparison of three hypotheses. *Educational Psychology*, *37*(4), 421–437. [CrossRef]

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. [CrossRef]

Holmes, J., & Gathercole, S. E. (2014). Taking working memory training from the laboratory into schools. *Educational Psychology*, *34*(4), 440–450. [CrossRef]

Ilkowska, M., & Engle, R. W. (2010). Trait and state differences in working memory capacity. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of individual differences in cognition: Attention, memory, and executive control* (pp. 295–320). Springer. [CrossRef]

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829–6833. [CrossRef]

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, *44*(2), 137–166. [CrossRef]

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*(4), 509–539. [CrossRef]

Kalyuga, S. (2009). *Managing cognitive load in adaptive multimedia learning*. IGI Global. [CrossRef]

Kalyuga, S., & Plass, J. L. (2017). Cognitive load as a local characteristic of cognitive processes: Implications for measurement approaches. In R. Z. Zheng (Ed.), *Cognitive load measurement and application: A theoretical framework for meaningful research and practice* (pp. 59–74). Routledge. [CrossRef]

Kim, Y.-S. G. (2016). Direct and mediated effects of language and cognitive skills on comprehension of oral narrative texts (listening comprehension) for children. *Journal of Experimental Child Psychology*, *141*, 101–120. [CrossRef]

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86. [CrossRef]

Kirschner, P. A., Sweller, J., Kirschner, F., & Zambrano, R. J. (2018). From cognitive load theory to collaborative cognitive load theory. *International Journal of Computer-Supported Collaborative Learning*, *13*(2), 213–233. [CrossRef] [PubMed]

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, *36*(5), 757–798. [CrossRef]

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289–325). Cambridge University Press. [CrossRef]

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*(4), 389–433. [CrossRef]

Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *Modern Language Journal*, *97*(3), 634–654. [CrossRef]

Limpo, T., & Alves, R. A. (2013). Modeling writing development: Contribution of transcription and self-regulation to Portuguese students' text generation quality. *Journal of Educational Psychology*, *105*(2), 401–413. [CrossRef]

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, *21*(4), 861–883. [CrossRef]

Logie, R. H., Camos, V., & Cowan, N. (Eds.). (2021). *Working memory: The state of the science*. Oxford University Press. [CrossRef]

Lusk, D. L., Evans, A. D., Jeffrey, T. R., Palmer, K. R., Wikstrom, C. S., & Doolittle, P. E. (2009). Multimedia learning and individual differences: Mediating the effects of working memory capacity with segmentation. *British Journal of Educational Technology*, *40*(4), 636–651. [CrossRef]

Martinussen, R., & Major, A. (2011). Working memory weaknesses in students with ADHD: Implications for instruction. *Theory into Practice*, *50*(1), 68–75. [CrossRef]

Mayer, R. E. (2017). Using multimedia for e-learning. *Journal of Computer Assisted Learning*, *33*(5), 403–423. [CrossRef]

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43–52. [CrossRef]

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer": Evidence from a meta-analytic review. *Perspectives on Psychological Science*, *11*(4), 512–534. [CrossRef]

Meyer, A., Rose, D. H., & Gordon, D. (2014). *Universal design for learning: Theory and practice*. CAST Professional Publishing.

Oberauer, K. (2019). Working memory and attention—A conceptual analysis and review. *Journal of Cognition*, *2*(1), 36. [CrossRef]

Olive, T. (2014). Toward a parallel and cascading model of the writing system: A review of research on writing processes coordination. *Journal of Writing Research*, *5*(2), 173–194. [CrossRef]

Paas, F., & van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*(1), 122–133. [CrossRef]

Paas, F., & van Merriënboer, J. J. G. (2020). Cognitive-load theory: Methods to manage working memory load in the learning of complex tasks. *Current Directions in Psychological Science*, *29*(4), 394–398. [CrossRef]

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*(2), 117–175. [CrossRef]

Pellicano, E. (2012). The development of executive function in autism. *Autism Research and Treatment*, *2012*, 146132. [CrossRef]

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., Dardick, W., & Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin*, *144*(1), 48–76. [CrossRef]

Peng, P., & Fuchs, D. (2016). A meta-analysis of working memory deficits in children with learning difficulties: Is there a difference between verbal domain and numerical domain? *Journal of Learning Disabilities*, *49*(1), 3–20. [CrossRef] [PubMed]

Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology*, *108*(4), 455–473. [CrossRef]

Peverly, S. T., Vekaria, P. C., Reddington, L. A., Sumowski, J. F., Johnson, K. R., & Ramsay, C. M. (2013). The relationship of handwriting speed, working memory, language comprehension and outlines to lecture note-taking and test-taking among college students. *Applied Cognitive Psychology*, *27*, 115–126. [CrossRef]

Preacher, K. J., & Sterba, S. K. (2018). Aptitude-by-treatment interactions in research on educational interventions. *Exceptional Children*, *85*(2), 248–264. [CrossRef]

Putnam, A. L., Sungkhasettee, V. W., & Roediger, H. L. (2016). Optimizing learning in college: Tips from cognitive psychology. *Perspectives on Psychological Science*, *11*(5), 652–660. [CrossRef]

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2016). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *32*(3), 164–171. [CrossRef]

Renkl, A. (2014a). The worked example principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 391–412). Cambridge University Press.

Renkl, A. (2014b). Toward an instructionally oriented theory of example-based learning. *Cognitive Science*, *38*(1), 1–37. [CrossRef]

Richland, L. E., & Burchinal, M. R. (2013). Early executive function predicts reasoning development. *Psychological Science*, *24*(1), 87–92. [CrossRef]

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science*, *316*(5828), 1128–1129. [CrossRef]

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. [CrossRef] [PubMed]

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher*, *39*(5), 406–412. [CrossRef]

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481–498. [CrossRef]

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research*, *64*(4), 479–530. [CrossRef]

Sala, G., & Gobet, F. (2020). Working memory training in typically developing children: A multilevel meta-analysis. *Psychonomic Bulletin & Review*, *27*(2), 423–434. [CrossRef]

Sana, F., Yan, V. X., & Kim, J. A. (2017). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology*, *109*(1), 84–98. [CrossRef]

Sana, F., Yan, V. X., Kim, J. A., Bjork, E. L., & Bjork, R. A. (2018). Does working memory capacity moderate the interleaving benefit? *Journal of Applied Research in Memory and Cognition*, *7*(3), 361–369. [CrossRef]

Sanchez, C. A., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition*, *34*(2), 344–355. [CrossRef]

Sanchez, C. A., & Wiley, J. (2009). To Scroll or not to scroll: Scrolling, working memory capacity, and comprehending complex texts. *Human Factors*, *51*(5), 730–738. [CrossRef]

Schmidt-Weigand, F., & Scheiter, K. (2011). The role of spatial descriptions in learning from multimedia. *Computers in Human Behavior*, *27*(1), 22–28. [CrossRef]

Schraw, G. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education*, *36*(1–2), 111–139. [CrossRef]

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science*, *11*(6), 771–799. [CrossRef] [PubMed]

Shipstead, Z., Lindsey, D. R. B., Marshall, R. L., & Engle, R. W. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language*, *72*, 116–141. [CrossRef]

Snyder, H. R., Miyake, A., & Hankin, B. L. (2015). Advancing understanding of executive function impairments and psychopathology: Bridging the gap between clinical and cognitive approaches. *Frontiers in Psychology*, *6*, 328. [CrossRef]

Suzuki, Y., Yokosawa, S., & Aline, D. (2020). The role of working memory in blocked and interleaved grammar practice: Proceduralization of L2 syntax. *Language Teaching Research*, *26*(4), 671–695. [CrossRef]

Swanson, H. L., Lussier, C. M., & Orosco, M. J. (2013). Cognitive strategies, working memory, and growth in word problem solving in children with math difficulties. *Journal of Learning Disabilities*, *48*(4), 339–358. [CrossRef] [PubMed]

Swanson, H. L., Orosco, M. J., & Kudo, M. (2016). Does growth in the executive system of working memory underlie growth in literacy for bilingual children with and without reading disabilities? *Journal of Learning Disabilities*, *50*(4), 386–407. [CrossRef]

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257–285. [CrossRef]

Sweller, J. (2017). *Cognitive load theory: Research that teachers really need to understand*. Centre for Education Statistics and Evaluation. Available online: https://education.nsw.gov.au/content/dam/main-education/about-us/educational-data/cese/2017-cognitive-load-theory.pdf (accessed on 10 May 2025).

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. [CrossRef]

Tarchi, C., Ruffini, C., & Pecini, C. (2024). Executive functions and multiple-text comprehension. *Learning and Individual Differences*, *108*, 102566. [CrossRef]

Tolmie, A. K., Ghazali, Z., & Morris, S. (2016). Children's science learning: A core skills approach. *British Journal of Educational Psychology*, *86*(3), 481–497. [CrossRef] [PubMed]

Tse, C. S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *Journal of Experimental Psychology: Applied*, *18*(3), 253–264. [CrossRef]

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132. [CrossRef] [PubMed]

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505. [CrossRef] [PubMed]

Unsworth, N., Miller, A. L., & Robison, M. K. (2021). Are individual differences in attention control related to working memory capacity? A latent variable mega-analysis. *Journal of Experimental Psychology: General*, *150*(7), 1332–1357. [CrossRef]

Unsworth, N., & Robison, M. K. (2020). Working memory capacity and sustained attention: A cognitive-energetic perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(1), 77–103. [CrossRef]

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2012a). Working memory capacity and retrieval from long-term memory: The role of controlled search. *Memory & Cognition*, *40*(2), 242–254. [CrossRef]

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2012b). Working memory capacity and retrieval limitations from long-term memory: An examination of differences in accessibility. *Quarterly Journal of Experimental Psychology*, *65*(12), 2397–2410. [CrossRef]

Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-student interaction: A decade of research. *Educational Psychology Review*, *22*(3), 271–296. [CrossRef]

Vanderberg, R., & Swanson, H. L. (2007). Which components of working memory are important in the writing process? *Reading and Writing: An Interdisciplinary Journal*, *20*(7), 721–752. [CrossRef]

Van Merriënboer, J. J. G., & Sweller, J. (2010). Cognitive load theory in health professional education: Design principles and strategies. *Medical Education*, *44*(1), 85–93. [CrossRef]

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*(1), 3–14. [CrossRef]

Vogel, S., & Schwabe, L. (2016). Learning and memory under stress: Implications for the classroom. *npj Science of Learning*, *1*(1), 16011. [CrossRef] [PubMed]

Wiley, J., & Jarosz, A. F. (2012). Working memory capacity, attentional focus, and problem solving. *Current Directions in Psychological Science*, *21*(4), 258–262. [CrossRef]

Xu, Z., Adam, K. C., Fang, X., & Vogel, E. K. (2018). The reliability and stability of visual working memory capacity. *Behavior Research Methods*, *50*(2), 576–588. [CrossRef]

Zhang, Y., Tolmie, A., & Gordon, R. (2023). The relationship between working memory and arithmetic in primary school children: A meta-analysis. *Brain Sciences*, *13*(1), 22. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Individual Differences in Student Learning: A Comparison Between the Student Approaches to Learning and Concept-Building Frameworks

Mark A. McDaniel [1,*], Christopher M. Wally [1], Regina F. Frey [2] and Hayley K. Bates [1]

[1] Center for Integrative Research on Cognition, Learning, and Education (CIRCLE), Washington University in St. Louis, St. Louis, MO 63130, USA; cwally@wustl.edu (C.M.W.); bates.h@wustl.edu (H.K.B.)
[2] Department of Chemistry, University of Utah, Salt Lake City, UT 84112, USA; gina.frey@utah.edu
* Correspondence: markmcdaniel@wustl.edu

Abstract

In cognitive science and education research, learning has been described to occur at surface and deep levels. Learners are thought to orient more toward one of these approaches to learning versus the other. In cognitive science, this has been assessed with a concept-building framework using objective function learning tasks to classify students as exemplar (surface) or abstraction (deep) learners. In education, the student approach to learning (SAL) framework has used self-report survey measures to classify learners as relying on surface approaches or deep approaches to learning. In two studies, we directly compared these two frameworks using self-report data from the Modified Approaches and Study Skills Inventory (M-ASSIST) and the Revised Study Process Questionnaire (R-SPQ-2F) along with objectively determined concept-building classifications from a computer-based function learning task. Potential links between exemplar learning and surface approaches and between abstraction learning and deep approaches were not found. We discuss possible explanations for the absence of empirical links, including inaccuracies in students' metacognitions regarding their learning, the measures, and possible differences between learning-content-dependencies of the survey responses versus content neutrality of the concept-building task. We conclude by suggesting directions for future work in assessing and comparing surface and deep learning across frameworks.

Keywords: student approaches to learning; concept-building; function learning; deep approach; surface approach; abstraction learner; exemplar learner; Modified Approaches and Study Skills Inventory (M-ASSIST); Revised Study Process Questionnaire (R-SPQ-2F)

1. Background

A central idea that spans theory and research in education and cognitive science is that learning occurs at multiple levels. One common distinction is between learning at a surface level versus learning at a deep level. In general, a surface level approach involves memorization of provided examples, whereas a deep level approach pushes for understanding of underlying concepts. This distinction can be seen in education within the widely used student approaches to learning (SAL) framework (Biggs, 1993; Entwistle & McCune, 2004). In cognitive science, levels of understanding can be described as varying from focusing on surface, literal features to deep understanding involving formation of mental models (Johnson-Laird, 1983; Kintsch, 1988; Mayer & Gallini, 1990). Similarly, in terms of conceptual representations, such representations can rely solely on memory of the training examples or on abstractions generated to capture the

relations of the training examples (e.g., Bourne, 1974; Koh & Meyer, 1991; D. R. Little et al., 2011; Medin & Schaffer, 1978; Nosofsky & Kruschke, 1992). Extending from this distinction is the idea that learners differ in their approaches to learning, with some learners adopting a relatively surface approach and others adopting a deeper approach. In this article, we review two frameworks that have attempted to establish and more completely characterize this posited individual difference in learners. The two frameworks have emerged in different spheres, one in the cognitive science domain and the other in the education domain, and with virtually no connections between the two made in the research literature. Yet, these frameworks share apparent similarities, prompting the question of whether they converge on the same individual difference, or at least have significant overlap. After describing the two frameworks, we report several novel studies to address this question.

1.1. Individual Differences in Learning Examples Versus Abstractions

Basic theoretical and empirical work from cognitive science has established that learning and understanding concepts can involve two fundamentally different representations. One kind of representation is based on learning the individual training examples (exemplar approach); the learned examples are then referenced when encountering new examples to categorize and draw inferences about the new examples (or problems) (e.g., Medin & Schaffer, 1978; Nosofsky & Kruschke, 1992). For instance, when learning about Lewis structure problems in general chemistry, a student could focus on memorizing the presented examples and solutions. These learned example problems and solutions are then relied upon to solve subsequent Lewis structure problems (e.g., when presented on a test). An example in biology could be when students learn about the concept of form follows function (which refers to the idea that the structure of a biological component (form) is adapted to perform a specific function). One could memorize different shape–function pairings of biological components and functions such as red blood cells providing oxygen or nerve cells (neurons) transmitting electrical–chemical signals. These relations could then be reproduced on a test.

Another kind of representation is based on an abstraction(s) from the trained examples that captures the critical features or basic principles that underlie the examples (Bourne, 1974; Koh & Meyer, 1991; D. R. Little et al., 2011). In this case, when learning about Lewis structure problems, a student would extract an understanding of the features and principles that underlie the particular solutions from the presented examples (e.g., that all non-hydrogen atoms would like a full shell of electrons, which is an octet, and that they obtain this octet by sharing electrons with another atom, either by forming a bond or by having lone pairs). These underlying principles are then applied to solve new Lewis structure problems (Frey et al., 2020). For the biology example, the student would use the illustrative examples—such as red blood cells—to abstract the concept of 'form follows function,' explaining why particular structures exist as they do and how those structures facilitate the functions they perform. For example, the biconcave shape of the red blood cells enables them to efficiently transport oxygen through narrow capillaries.

The key assumption of this framework is that individual learners can differ in terms of which approach they adopt when confronted with a learning challenge. In a range of conceptual tasks—category learning (J. L. Little & McDaniel, 2015), function learning (McDaniel et al., 2014), multiple-cue prediction learning (Hoffmann et al., 2014; Juslin et al., 2003), and skill learning (Bourne et al., 2010)—recent evidence supports this major tenet that an individual learner relies predominantly on either an exemplar or an abstraction approach to learn a particular conceptual task. Further, in a laboratory study, learners completed function-learning tasks and unrelated categorization tasks over the course of several weeks (McDaniel et al., 2014). Based on their extrapolation performance in the function-learning task, learners were identified as having oriented toward memorizing the particular training

pairs (each input value/output value pair) or as attempting to abstract the underlying functional relations between inputs and outputs (a bilinear "V-shape"). Importantly, learners' approaches on the function-learning task significantly predicted learners' performances on subsequent category-learning tasks, such that those who displayed exemplar learning on the function-learning task showed example-driven behavior (on categorization transfer tests) consistent with an exemplar representation, whereas those learners who displayed abstraction concept-building also showed abstraction-driven categorization behavior. Thus, individual differences in concept-building tendencies tended to persist across very different learning tasks, even with those tasks presented several weeks apart. Converging with these findings, Herzog and Goldwater (in press) found that participants consistently displayed either an exemplar-learning approach or an abstraction approach across five distinct laboratory tasks (drawn from function learning, category learning, associative learning, and analogy).

Most critically for the present project, and for science learning in general, we have recently shown in several studies that the individual differences in concept-building tendencies, as indexed by the laboratory learning task from McDaniel et al. (2014), extend to the learning challenges in complex STEM topics (Frey et al., 2017, 2020; McDaniel et al., 2022, 2018). Students in chemistry classes and biology classes who displayed exemplar concept-building tendencies on the function-learning task (exemplar learners) tended to focus on learning particular example problems and solutions that are presented in class and homework (e.g., for Lewis structure problems, as mentioned above (Frey et al., 2020)). By contrast, students who abstracted in the function-learning task (abstraction learners) attempted to abstract the principles and concepts that underlie the class and homework problems (e.g., for Lewis structure problems). In terms of exam performances, the exemplar learners and abstraction learners, as identified on the basic function-learning task, performed equally well on course-exam problems that were similar to those presented in class (reproduction questions), but the exemplar learners performed poorly relative to the abstraction learners on exam items that were not similar to class problems, i.e., those that required generalization of the topic (transfer questions) (McDaniel et al., 2018, 2022). Critically, based on assessments of general ability, we were able to conclude that this divergence across students on transfer exam items did not simply reflect a difference in general ability or intelligence.

1.2. The Student Approaches to Learning Framework

A complementary framework often used in educational spheres is the student approaches to learning, SAL (Biggs, 1993; Biggs et al., 2001; Entwistle & McCune, 2004). The central idea is that some students adopt a deep approach to learning with the intention to understand the content, whereas other students adopt a surface learning approach with the intention to reproduce the content (Marton & Säljö, 1976). In this framework, this distinction emphasizes that students have different intentions when confronting a learning task (such as studying), and that these intentions in turn influence the strategies they use (see (Google et al., 2023), for applications of this framework to STEM learning). This categorical combination of intention and strategies (deep vs. surface) was termed "approaches to learning" (Marton & Säljö, 1997). For example, a key observation was that when students were given the same learning task (i.e., reading an academic article to prepare for answering questions about it), students varied in their descriptions of how they went about accomplishing the task (Marton & Säljö, 1997). The descriptions revealed a surface approach for some and a deep approach for others.

A student's approach to learning is measured using a Likert-scale survey. These surveys were initially developed by a two-step methodology: qualitative analyses of students' descriptions about their study processes followed by quantitative item creation. Many surveys were designed for generic learning contexts such as the Approaches and Study Skills Inventory for

Students (ASSIST) (Tait et al., 1997), the Study Process Questionnaire (SPQ) (Biggs, 1987a, 1987b), and the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich et al., 1991). Other surveys were developed from these original surveys, such as the revised two-factor Study Process Questionnaire (R-SPQ-2F) (Biggs et al., 2001), or modified for discipline-specific learning contexts, such as the Modified Approaches and Study Skills Inventory (M-ASSIST) in chemistry (Bunce et al., 2017; Malinakova, 2024). Findings on the relationship between learning approaches and student outcomes are mixed. Some studies show a positive correlation between a deep learning approach and student outcomes (e.g., Zeegers, 2001; May et al., 2012; Bunce et al., 2017), whereas others report no significant correlation between a deep-learning approach and outcomes (e.g., dos Santos Rodrigues & Gomes, 2020; Rowe, 2002; Maciejewski & Merchant, 2016).

1.3. The Present Studies

In considering the concept-building tendency and SAL frameworks that have been developed from different theoretical bases, research, and disciplines, a natural question is whether the two are converging on a similar idea to characterize individual differences in learning (cf. Goldwater et al., 2018, Experiment 4). Favoring this possibility is the theoretical premise that pushing for understanding (a deep approach in SAL) is a process that can involve abstracting the underlying principles or regularities of the target content (abstraction learners in the concept-building framework). By the same token, adopting a surface approach emphasizes reproduction of the target content (the SAL view); reproduction, at least in part, requires committing to memory the specific examples provided in the target content (exemplar learners in the concept-building framework).

To investigate the possible overlap between these two fundamental approaches concerning individual differences in learning, we conducted two related, novel studies. The studies adopted the same methodology of requiring a group of college students to complete both the self-report surveys used in the education literature based on the SAL framework and the function-learning task developed in the cognitive science learning literature. It warrants mention that the SAL instruments provide two indices for each respondent, a score reflecting their tendency toward a deep approach to learning and a score reflecting their tendency toward a surface learning approach; in contrast, the laboratory (function) learning task provides a categorical index indicating whether the respondent tends to use an abstraction approach or an exemplar approach to learning.

In these studies, we used correlational analyses and inferential statistics to test two competing hypotheses. Because the literature (to our knowledge) has not directly compared the SAL and concept-building frameworks, we did not have clear expectations for how they might be related. We thought it equally plausible that there may be positive associations between deep and abstraction learning and surface and exemplar learning, as described above, or that these approaches might not be related at all. Given this, we tested the following two hypotheses:

Hypothesis 1. *Learners who adopt an abstraction approach to concept learning will report more extensive reliance on a deep approach to their learning than will learners who adopt an exemplar approach. On the other hand, learners who adopt an exemplar approach to concept learning will report more extensive reliance on a surface approach to their learning than will learners who adopt an abstraction approach.*

This hypothesis originates from the possible theoretical linkages that hinge on the idea that there is a close correspondence between learning processes and the consequent representations of what is learned. Surface learning processes (per SAL, committing items to memory) can be tied to literal representations that capture the items in the instructional

material (e.g., for an instructional text, the representation would contain the words and sentences as presented in the text (Kintsch, 1988; Kintsch & van Dijk, 1978)). Similarly, when viewing a sample of values capturing a functional relation between two variables (e.g., amount of precipitation and crop growth), the representation would be those specific example values (DeLosh et al., 1997; McDaniel et al., 2014). Deep learning processes (per SAL, trying to gain understanding) are theoretically tied to representations that capture relations and abstractions among items. For instance, in an instructional text, it would be a mental model of what the text conveys (Kintsch, 1988; Mayer & Gallini, 1990); whereas, for presentation of function values, it would be an abstraction of the functional relationship conveyed by the training values (McDaniel et al., 2014).

Based on this hypothesis (and noting that our analyses coded an abstraction learner as "0" and an exemplar learner as "1"), there should be a significant *negative* correlation between concept-building tendency and the magnitude of endorsing a deep approach to learning, and there should be a significant *positive* correlation between concept-building tendency and the magnitude of endorsing a surface approach to learning. A related pattern that would solidify the merit of this hypothesis is that in a two-factor analysis of variance (ANOVA) with concept-building tendency and reported approaches to learning as the independent variables, there should be a significant cross-over interaction such that abstraction learners display a higher deep-approach to learning score than exemplar learners, whereas exemplar learners display a higher surface-approach to learning score than abstraction learners.

Hypothesis 2. *The particular concept-building approach adopted by a learner has no discernible relationship to the self-reported approaches to learning endorsed by a learner.*

This possibility may be equally plausible and stems from important distinctions between the two frameworks. For example, SAL emphasizes a student's *intention* in approaching a learning task (Marton & Säljö, 1976). Learners may not be able to achieve their intentions, however, especially in terms of trying to achieve understanding (e.g., Gouravajhala & McDaniel, 2025). For instance, in approaching a category learning task, some learners initially reported attempting to learn a rule (abstraction) that captured the range of training examples within a category (Gouravajhala et al., 2020). Despite their intention to learn a rule, by the end of training, a subset of these learners displayed an exemplar representation; i.e., their learned representations (for the categories) did not support a rule-based pattern of classifying novel category examples. The concept-building framework emphasizes, by contrast, the cognitive representations that learners actually acquire during learning—underlying abstractions (principles) that capture a concept or surface representations based on memorizing examples; it is these representations that ultimately reflect what the student is learning and index the degree to which individual learners achieve relatively deep learning.

Moreover, the SAL framework has relied on student self-reports of their learning approaches, thereby depending on the accuracy of students' metacognitive awareness of their learning. Yet, the basic cognitive science literature has shown that students' judgments of learning and of comprehension can be remarkably poor (Dunlosky & Metcalfe, 2008). Similarly, in a classroom chemistry study, a number of students judged that they had a good understanding of the principle even though think-aloud protocols revealed they did not (Frey et al., 2020). Therefore, poor metacognition—for instance when students self-report taking a deep approach to learning—may cloud the validity of classifications of learning approaches as deep or surface. For this reason or for the reason noted in the preceding paragraph (or both), there may be a negligible association between gauging students' approaches to learning using common instruments inspired by SAL (introduced in the

studies reported below) and the objective learning task inspired by the concept-building approach framework.

Repeatedly observing correlations that do not differ from "0" and finding no significant cross-over interaction in the ANOVAs (in each study) would be strongly consistent with this hypothesis.

2. Studies 1a and 1b

The SAL framework encompasses several complexities. One group of theorists assumes that students adapt their approaches to learning depending on the learning context, influenced by individual factors (e.g., student perceptions of the learning environment that are shaped by prior experience) and course-related factors (i.e., course design, the teaching and learning environment, and assessment procedures). For example, students are more likely to utilize surface approaches when faced with a learning environment that rewards rote memorization of facts. Further, the SAL framework has been developed in different discipline-specific learning contexts such as biology (dos Santos Rodrigues & Gomes, 2020; Nelson Laird et al., 2008; Lee et al., 2016), chemistry (Li et al., 2013; Sinapuelas & Stacy, 2015; Bunce et al., 2017; Atieh et al., 2020), mathematics (Maciejewski & Merchant, 2016; Lahdenperä et al., 2019), physics (Prosser et al., 2000; Selcuk, 2010), and engineering (Rowe, 2002; Zakariya et al., 2023). Consequently, the specific study strategies defined as deep and surface approaches to learning are assumed by some SAL proponents to vary across disciplines and learning environments (e.g., Sinapuelas & Stacy, 2015; Case & Marshall, 2004).

To capture the context dependency of the learning environment for determining in part a student's approach to learning, Study 1a included an examination of the relationship between a student's concept-building approach and their student-learning approach in a general chemistry classroom. Students in this class completed both the concept-building task and the Modified Approaches and Study Skills Inventory (M-ASSIST) survey, thereby enabling a direct classroom comparison of whether the deep and surface approaches in the SAL framework capture a similar characterization of individual learning differences based on the cognitive representations students construct during a learning task (measured by the concept-building task). The M-ASSIST is a shorter adaptation of the ASSIST survey[1] developed specifically for chemistry and has been used in both general chemistry and organic chemistry (Bunce et al., 2017; Malinakova, 2022, 2024). The M-ASSIST consists of 12 items focusing on the deep approach (6 items) and the surface approach (6 items). The M-ASSIST was selected for the present studies because of its adaptation for chemistry courses and its focus on deep and surface approaches of the SAL framework.

Study 1b was motivated by an additional complexity in the SAL literature. Some theorists suggest that the two learning approaches—deep and surface—reflect overarching characteristics that transcend the particular learning context (Entwistle et al., 1979; Entwistle & McCune, 2004). With the deep approach, students are intrinsically motivated to understand the intentions of the educational tasks, to connect the task to their own real-world experiences, to integrate aspects of the tasks with prior knowledge, and to ask questions about the tasks. By contrast, with the surface approach, the students are driven by extrinsic factors, such as fear of failure, which lead to low-level cognitive activities like rote learning or selective memorizing without understanding. Additionally, they invest minimal effort to understand the task's purpose, even when high-level cognitive engagement is required.

To the extent that the SAL deep and surface approaches to learning are overarching characteristics of the learner, then those characteristics might best be evidenced outside of a particular classroom context. To test this possibility, in Study 1b, we administered the M-ASSIST to students in a laboratory setting and compared these responses to their concept-building tendencies (determined in the same laboratory setting).

Hypothesis 1 (from the general introduction) would be strongly supported if at least one of these two studies (Study 1a or Study 1b) demonstrated significant associations between the M-ASSIST deep scale and a learners' tendency toward an abstraction concept-building approach and between the M-ASSIST surface scale and learners' tendency toward an exemplar-based concept-building approach. The absence of associations between the M-ASSIST survey responses and learners' concept-building tendencies in both Studies 1a (reflecting the possibility of an assumption of course specific influences of students' deep vs. surface approaches to learning) and 1b (reflecting an overarching tendency outside the context of a specific course) would support Hypothesis 2.

2.1. Method

2.1.1. Participants

Study 1a examined students enrolled in a first-semester general chemistry course at a small health professions school in the midwestern United States in fall 2017. Note that the concept-building data for this study were partially reported in previous work (Frey et al., 2020), but the data related to the M-ASSIST was not analyzed and is reported here for the first time. The project was approved by the institutional review boards at the university where the data were collected (IRB ID no. 2017-39) and the university where the data were analyzed (IRB ID no. 201710100). Consent was provided by 108 out of 123 (87.8%) of these students. Study 1b included 77 students electing to participate in a research laboratory experiment as one of several options to fulfill a research credit requirement for their psychology courses at a mid-sized private research university in the midwestern United States in fall 2024. The project was approved by the university's institutional review board (IRB ID no. 202308131).

2.1.2. Concept-Building Task

Participants' concept-building tendency (exemplar or abstraction learners) was determined using the same concept-building task established in previous work (McDaniel et al., 2014, 2018; Frey et al., 2017, 2020). Participants had no prior knowledge or exposure to the task, as it involved a fictional scenario with an alien organism and two non-existent elements. The task was completed on a computer, and participants were asked to imagine they had been hired by NASA to study a new organism found on Mars that absorbs an element called Zebon and releases an element called Beros. The goal of the task was to predict an output variable (quantity of Beros) based on the input variable (quantity of Zebon). Unknown to the participants, the input-output points followed either an inverted-V function (Study 1a) or a regular-V function (Study 1b).

Participants first completed a training phase, in which output predictions were made based on given input values, followed by feedback showing the correct value and the amount of error in their prediction. Figure 1 shows an example of a training trial with input, output, and feedback screenshots. In total, there were 20 unique input values (all odd numbers between 61 and 99 for Study 1a; all odd numbers between 81 and 119 for Study 1b), which were presented once each during a training block. Participants completed a total of 10 training blocks (200 individual trials), with the order of input values varying across each training block. No instructions were given as to how long to spend on each trial; the participants paced themselves throughout the task. Following each block, participants were shown their mean prediction error (mean absolute error, MAE) for that block. Additionally, after block two, participants saw their previous MAE and a message indicating whether they had reduced or failed to reduce their error ("Your accuracy IMPROVED. Keep up the good work!" vs. "Your accuracy DID NOT IMPROVE. Keep working to improve your predictions!"). After 10 training blocks, if the participant achieved a MAE <10, training concluded. In Study 1a, if participants had not met the threshold, up to three additional

training blocks were presented, with training ending if MAE dropped below 10 on either block 11 or block 12. Training ended after block 13 (trial 260) for everyone in this version, regardless of whether the learning criterion was met. In Study 1b, training ended following block 10 for everyone.

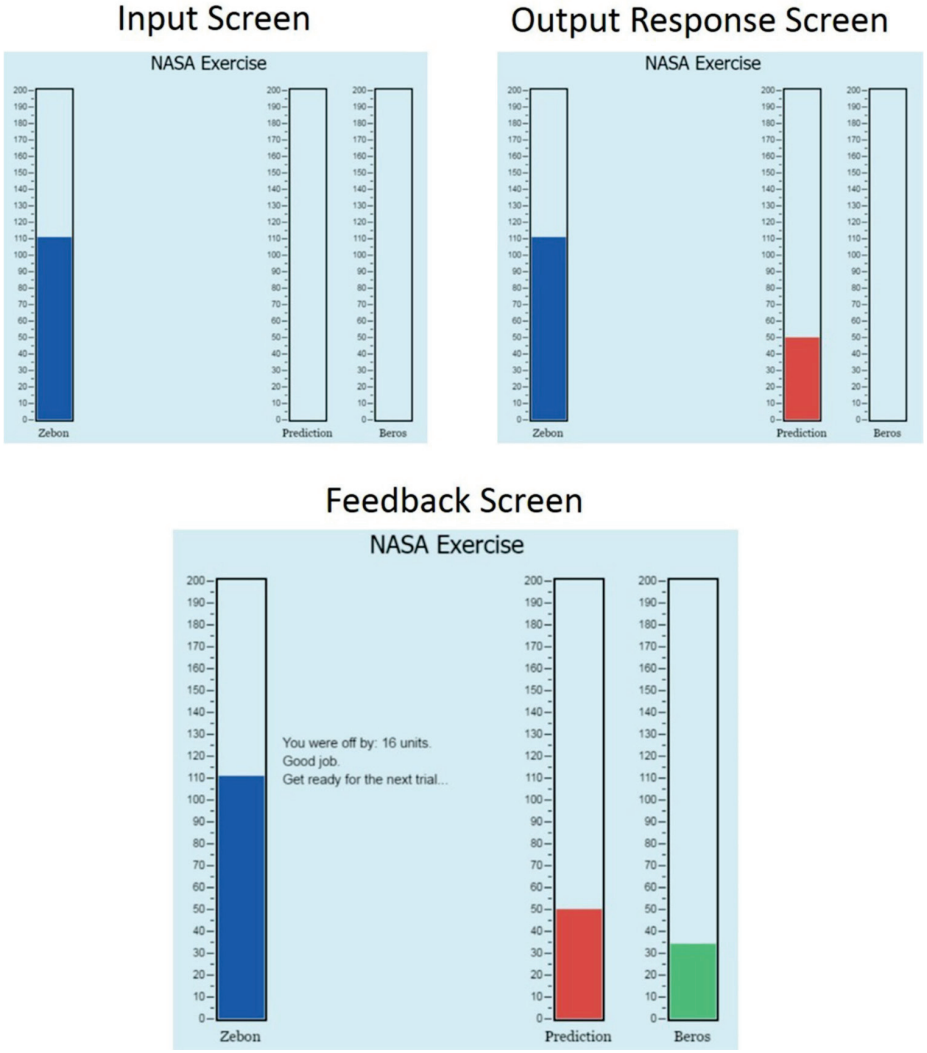


Figure 1. Screenshots from the concept-building task. ***Note.*** Adapted from (McDaniel et al., 2014). Copyright 2014 by the American Psychological Association.

Following the training blocks, participants completed a 36-trial transfer phase in which they predicted output values for novel (previously unseen) inputs. The procedure was similar except that no feedback was given after each trial. After making a prediction, participants saw the message "Prediction recorded. Get ready for the next trial." The transfer phase featured 30 extrapolation trials, composed of odd numbers outside the training range (all odd numbers between 31 and 59 and between 101 and 129 for Study 1a; all odd numbers between 51 and 79 and between 121 and 149 for Study 1b); it also included six interpolation trials, featuring even numbers within the training range (64, 72, 80, 88, 94, and 100 for Study 1a; 84, 92, 100, 108, 114, and 120 for Study 1b). The extrapolation and interpolation trials were intermixed for the transfer phase. Participants took around 40 min on average to complete the full task.

As with the previous studies using the concept-building task, learners were classified using a two-step process. First, participants with a final training block MAE greater than or equal to 10 were classified as non-learners. The rest of the participants falling

below the cutoff were considered "learners," and were then further classified using their extrapolation MAEs from the transfer trials. By using a simple exemplar model as a comparison, one would predict a flat extrapolation extending from the edges of the training domain (represented by dashed horizontal lines in Figure 2). The functions used would result in a simple exemplar model making a prediction of 148 for every extrapolation trial in Study 1a and a prediction of 52 in Study 1b, leading to a MAE of 34.72 in both cases. Note that any set of predictions that average 148 or 52, respectively, and never overestimate the output value produce the same MAE of 34.72. The extrapolation MAE and surrounding 95% confidence interval were calculated for each learner.

If the upper-bound limit of the 95% confidence interval was below 34.72 (i.e., the learner's predictions were statistically significantly better than a simple exemplar model), the learner was classified as an abstraction learner. This assumes that to outperform an exemplar model, learners must have extracted some information about the function rule during the training trials that was then applied during extrapolation. The remaining learners whose upper-bound extrapolation MAE values were greater than 34.72 (i.e., did not outperform a simple exemplar model) were classified as exemplar learners. The assumption is that these participants learned specific input-output associations but did not extract the function rule necessary to make accurate predictions on the novel transfer (extrapolation) inputs.

To represent these distinctions visually, Figure 2 displays the mean output predictions for each input value by abstraction and exemplar learners in Study 1a and Study 1b. Inspection of Figure 2 indicates that accuracies of predictions for abstraction and exemplar learners in the training range (interpolation) are both high (i.e., all learners look the same in the training range). In contrast, it can be seen that accuracy of extrapolation trials diverge, such that abstraction learners make predictions that follow the given function and exemplar learners do not extrapolate much beyond the value for the most extreme training point[2] (i.e., tends to be relatively flat in predictions hovering around the value of the closest training point).

2.1.3. The Modified Approaches and Study Skills Inventory (M-ASSIST)

The M-ASSIST is a 12-item measure developed by Bunce et al. (2017) to assess students' approaches to studying (Bunce et al., 2017; see Supplementary Materials for full set of survey items). It is a brief, modified version of the Approaches and Study Skills Inventory for Students (ASSIST) (Tait et al., 1997). The M-ASSIST contains six items comprising a deep scale, reflecting deep approaches and strategies to learning, and six items comprising a surface scale, reflecting more superficial approaches and concerns about learning. Response options range from 'disagree' to 'agree' on a five-point Likert scale. Deep approach and surface approach scales were scored by averaging the individual item ratings across each scale. The measure was found to be a good fit for a two-factor structure across two semesters of data collected by Bunce et al. (2017) with model fit indices: CFI = 0.95, RMSEA = 0.05, and SRMR = 0.05 for the 1st semester, and CFI = 0.96, RMSEA = 0.04, and SRMR = 0.04 for the 2nd semester. Standardized factor loadings ranged from 0.36 to 0.70 for the deep items, and 0.43 to 0.82 for the surface items, with all p's < 0.05.

2.1.4. Procedure

For Study 1a, all students enrolled in the chemistry course completed the concept-building task following their first course exam during a supervised laboratory period approximately 5–6 weeks into the course. Students were given one-hour to complete the concept-building task. The M-ASSIST was completed 2–4 weeks later via an online survey at the student's leisure during this time period. Only consenting students' data were included in the analyses. Of the 108 consenting students, 54 (50%) were classified as abstraction learners, 28 (25.9%) exemplar learners, 25 (23.1%) non-learners, and 1 did not complete the task.

For Study 1b, participants were directed to a private room with a computer and proceeded to complete the concept-building task. Once the task was finished for participants, they were given five online surveys to complete, with the M-ASSIST included as the fourth survey. Participants were given up to 90 min to complete the concept-building task and surveys. Of the 77 consenting participants, 30 (39%) were classified as abstraction learners, 31 (40%) as exemplar learners, and 16 (21%) as non-learners.

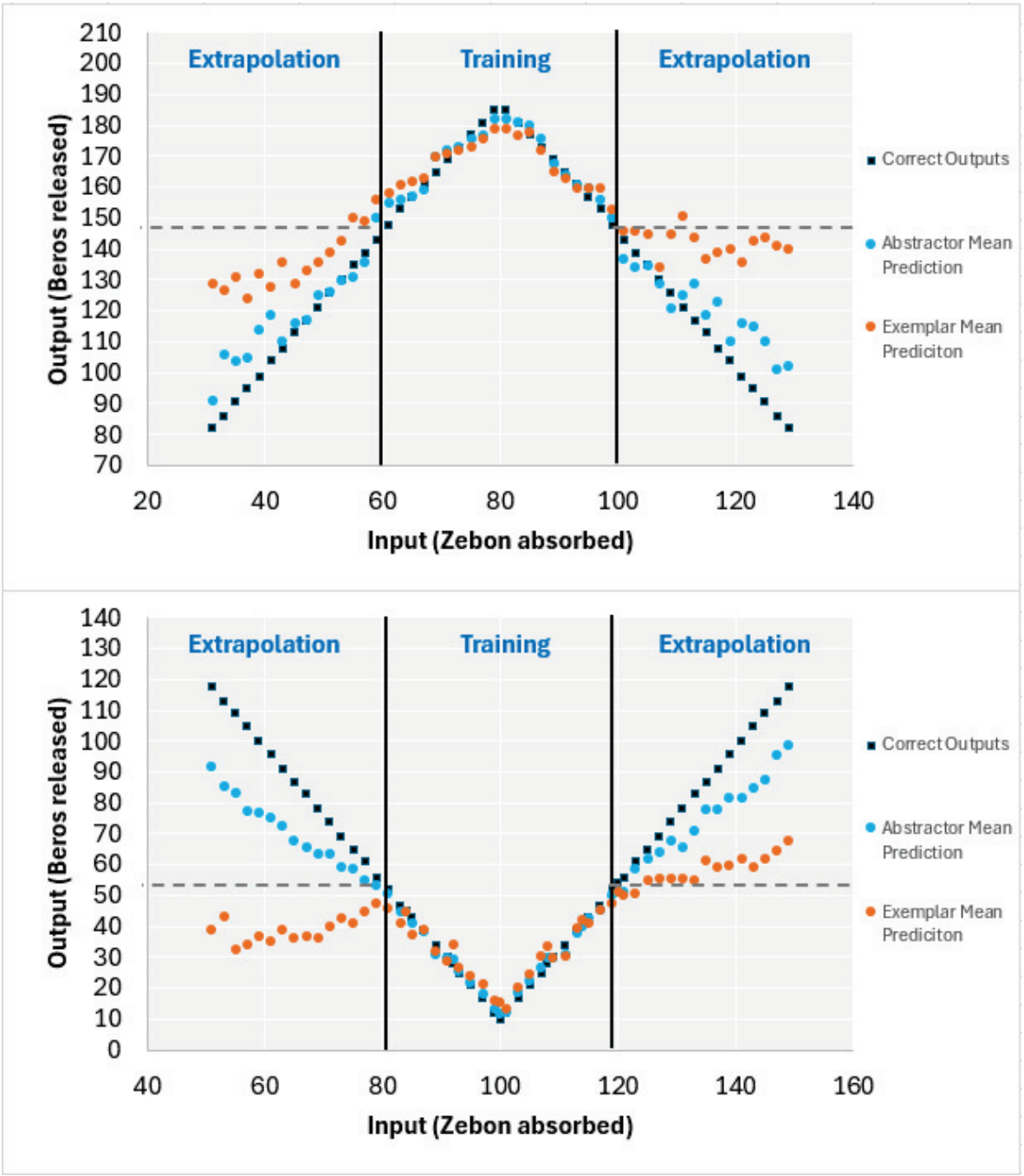


Figure 2. Input and output values from final training block and transfer trials for participants in Study 1a (**Top Chart**) and participants in Study 1b (**Bottom Chart**). ***Note.*** Vertical black lines represent the boundaries of the training range. Dashed horizontal lines represent theoretical exemplar model predictions based on learning the outputs at the boundaries of the training range and predicting those same outputs during extrapolation. Black squares represent the correct output values, based on the function. The blue circles represent the mean prediction across all abstraction learners for each input value. The orange circles represent the mean prediction across all exemplar learners for each input value.

2.1.5. Analyses

All statistical analyses were conducted using SPSS version 29. To examine associations between deep approach and surface approach strategies on the M-ASSIST with abstraction and exemplar concept-building approaches, correlation coefficients were calculated. Concept-building tendency was dichotomously coded as 0 for abstraction learners and 1 for exemplar learners, and separate point-biserial correlations were computed with the deep approach and surface approach scores (treated as continuous variables). To increase sensitivity, we also included the continuous variable extrapolation mean absolute error (MAE) to index concept-building tendency and compared this with the deep and surface approach scores. A 2 (Concept-Building Approach: Abstractor vs. Exemplar; between-subjects) X 2 (M-ASSIST: Deep Approach vs. Surface Approach; within-subjects) mixed factor analysis of variance (ANOVA) was used to evaluate if abstraction learners endorsed deep approach items more strongly than surface approach items, and exemplar learners endorsed surface items more than deep items (an interaction effect). Statistical significance was set at $p < 0.05$ for all tests. As a general guideline for effect size estimates, Pearson's r values of 0.1 are considered small, 0.3 medium, and 0.5 or higher large; partial-eta squared values (associated with the ANOVA effects) of 0.01 are considered small, 0.06 medium, and 0.14 large (Cohen, 1988).

Power analyses using G*Power 3.1.9.7 (Faul et al., 2007, 2009) were conducted to determine the power available to detect medium-size and large-size effects for Study 1a and Study 1b. At a significance level of $p = 0.05$, Study 1a ($N = 82$) provided 79% power to detect a medium correlation (99% power for large correlation) and 87% power to detect a medium effect size (99% power for large effect size) when testing the interaction between concept-building and learning approach using the M-ASSIST. Study 1b ($N = 61$) provided 66% power to detect a medium correlation (99% power for large correlation) and 73% power to detect a medium effect size (98% power for large effect) for the interaction. To provide greater power to detect associations between concept-building and student approaches to learning, data from Study 1a and Study 1b were pooled in a subsequent analysis. At a significance level of $p = 0.05$, this combined sample ($N = 143$) provided 96% power to detect a medium correlation and 98% power to detect a medium effect size.

2.2. Results and Discussion

Table 1 displays correlation coefficients for concept-building approach (as point biserial correlations), the continuous index used to determine learner category (extrapolation MAE), M-ASSIST deep approach scores, and M-ASSIST surface approach scores for Studies 1a and 1b. Concept-building approach was significantly positively correlated with the deep approach responses of the M-ASSIST in Study 1a, but not in Study 1b. The Study 1a correlation indicates that exemplar learners were more likely to endorse a greater number of deep approach items than abstraction learners—counter to Hypothesis 1. Concept-building approach was not correlated with the surface approach responses of the M-ASSIST in either study. Further, extrapolation MAE was not significantly related with deep or surface approach scores in either study. Finally, the deep and surface approach scales of the M-ASSIST were not significantly correlated with one another in either study.

Figure 3 displays the average scores for the deep approach and surface approach scales of the M-ASSIST for abstraction and exemplar learners in Studies 1a and 1b. For Study 1a, all learners endorsed significantly more deep approach items than surface approach items, $F(1, 80) = 53.47$, $p < 0.001$, $\eta^2_p = 0.401$. Unique to Study 1a, exemplar learners endorsed items more strongly overall than abstraction learners, $F(1, 80) = 5.56$, $p < 0.001$, $\eta^2_p = 0.065$. Concept-building approach did not significantly interact with M-ASSIST learning approach, $F(1, 80) = 0.68$, $p = 0.411$, $\eta^2_p = 0.008$.

For Study 1b, as with Study 1a, all learners endorsed significantly more deep approach items than surface approach items, $F(1, 59) = 33.88$, $p < 0.001$, $\eta^2_p = 0.365$. However, in Study 1b exemplar learners and abstraction learners did not significantly differ in their endorsement of items, $F(1, 59) = 0.20$, $p = 0.66$, $\eta^2_p = 0.003$. Most importantly, concept-building tendency again did not significantly interact with M-ASSIST learning approach responses, $F(1, 59) = 0.30$, $p = 0.585$, $\eta^2_p = 0.005$.

Table 1. Correlations for concept-building approach and M-ASSIST deep approach and surface approach scales in Study 1a (upper table section) and Study 1b (lower table section).

Study 1a Variable	1	2	3	4
1. Concept-Building Approach [a]	-			
2. Extrapolation MAE	0.83 *	-		
3. M-ASSIST Deep Approach	0.29 *	0.19	-	
4. M-ASSIST Surface Approach	0.09	0.15	−0.08	-
Study 1b Variable	**1**	**2**	**3**	**4**
1. Concept-Building Approach [a]	-			
2. Extrapolation MAE	0.87 *	-		
3. M-ASSIST Deep Approach	−0.10	−0.06	-	
4. M-ASSIST Surface Approach	0.01	0.04	−0.13	-

Note.[a] 0 = abstraction learners, 1 = exemplar learners. * $p < 0.05$. $n = 82$ for Study 1a correlations, $n = 61$ for Study 1b correlations.

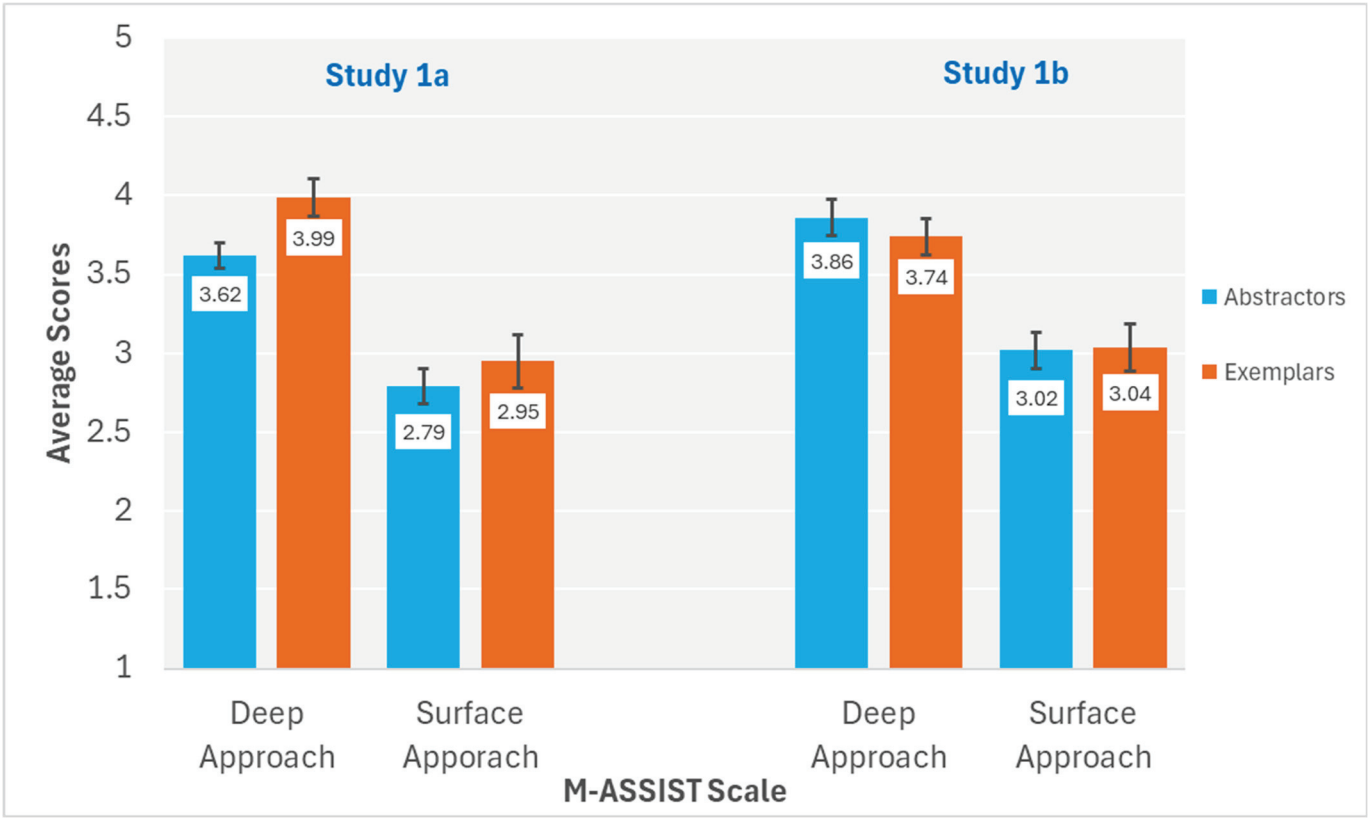


Figure 3. M-ASSIST deep approach and surface approach average scores for abstractor and exemplar learners in Studies 1a and 1b. ***Note.*** Error bars represent standard error of the mean. In Study 1a, there were 54 abstractors and 28 exemplars. In Study 1b, there were 30 abstractors and 31 exemplars.

Combined Analyses

Table 2 displays correlation coefficients for concept-building approach, extrapolation MAE, M-ASSIST deep approach, and M-ASSIST surface approach scores for the combined

data of Studies 1a and 1b. Concept-building approach was not significantly related with the deep approach responses or with the surface approach responses of the M-ASSIST. Similarly, extrapolation MAE was not significantly correlated with deep or surface approach scores. Additionally, the deep and surface approach scales of the M-ASSIST were not significantly correlated with one another.

Table 2. Correlations for concept-building approach and M-ASSIST deep approach and surface approach scales in combined Study 1a and 1b sample.

Variable	1	2	3	4
1. Concept-Building Approach [a]	-			
2. Extrapolation MAE	0.83 *	-		
3. M-ASSIST Deep Approach	0.12	0.05	-	
4. M-ASSIST Surface Approach	0.08	0.09	−0.09	-

Note.[a] 0 = abstraction learners, 1 = exemplar learners. * $p < 0.05$. $n = 143$ for combined sample correlations.

Figure 4 displays the average scores for the deep approach and surface approach scales of the M-ASSIST for abstraction and exemplar learners for the pooled data. Most importantly, converging with the individual analyses of Studies 1a and 1b, there was no hint of an interaction between concept-building tendency and self-reported learning approaches on the M-ASSIST, $F(1, 141) = 0.02$, $p = 0.885$, $\eta^2_p = 0.000$. Also as before, significantly more deep approach items were endorsed by learners compared to surface approach items, $F(1, 141) = 87.05$, $p < 0.001$, $\eta^2_p = 0.382$. With the pooled data, the tendency seen in Study 1a for exemplar learners to more strongly endorse M-ASSIST items regardless of approach (deep or surface) relative to abstraction learners was no longer significant, $F(1, 141) = 2.96$, $p = 0.088$, $\eta^2_p = 0.021$.

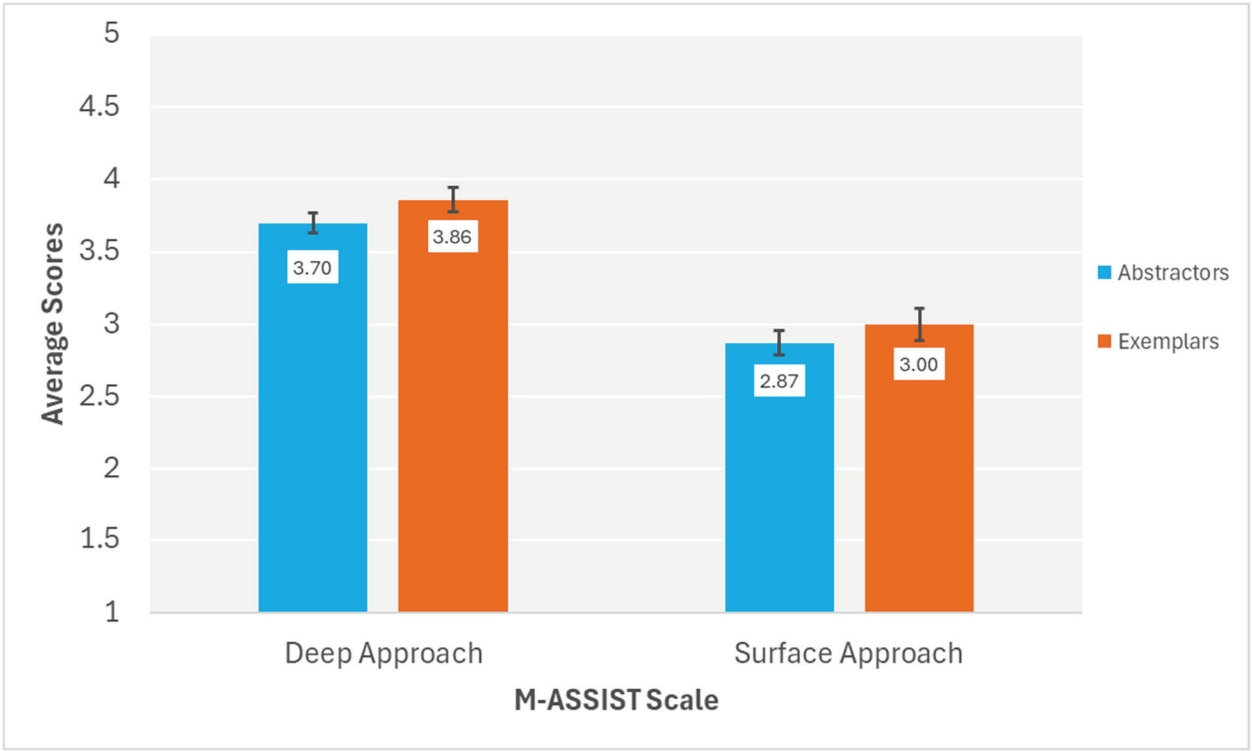


Figure 4. M-ASSIST deep approach and surface approach average scores for abstractor and exemplar learners for combined data of Studies 1a and 1b. ***Note.*** Error bars represent standard error of the mean. In the combined sample, there were 84 abstractors and 59 exemplars.

Hence, overall, when the results were considered across the general chemistry classroom and the laboratory settings, the patterns were straightforward. There were no significant correlations between the deep approach score from M-ASSIST and concept-building tendency or between the surface approach score from M-ASSIST and concept-building tendency. Reinforcing these correlation results, there was no crossover interaction in the ANOVAs between abstraction learning and deep approach scores or exemplar learning and surface approach scores.

Thus, these findings provide evidence for Hypothesis 2, which states that the concept-building tendency of a learner is not related to the learner's self-reported approaches to learning. (Study 1a produced the anomalous finding that exemplar learners endorsed a greater number of deep approach items than abstraction learners, but this result did not stand in Study 1b or when combining the datasets. Perhaps, this was due, in part, to participants in Study 1a completing the concept-building task and M-ASSIST several weeks apart, during which some learners' approaches to learning may have shifted. To avoid this possibility, Study 1b required participants to complete both assessments during the same session.) To provide further support for Hypothesis 2 of a non-relationship between these factors, a Bayesian alternative to traditional null-hypothesis significance testing was performed to test for an interaction (examined in the ANOVA) following the procedure outlined by Masson (2011) and originally formulated by Wagenmakers (2007). A Bayes factor value was computed (from the delta Bayesian Information Criterion) to determine the posterior probabilities of the null hypothesis model. Study 1a yielded a posterior probability of 0.865 for the null hypothesis, suggesting "positive" evidence for the null hypothesis model (i.e., no interaction) (following guidelines from (Raftery, 1995)). Study 1b yielded a posterior probability of 0.870 for the null hypothesis, again providing "positive" evidence for the null. The combined data yielded a posterior probability of 0.922 for the null hypothesis model, showing stronger positive evidence that concept-building tendencies were not related to students' self-reports of deep or surface approaches to learning.

3. Study 2

Studies 1a and 1b relied on the M-ASSIST to determine students' approaches to learning, which was designed specifically for use in chemistry courses. To better generalize outside of a chemistry learning context, in Study 2 we administered another commonly used SAL inventory, the two-factor Revised Study Process Questionnaire (R-SPQ-2F) to determine student approaches to learning (Biggs et al., 2001). In addition to being designed for generic learning contexts, the R-SPQ-2F has been used in mathematics (Maciejewski & Merchant, 2016), physics (Prosser et al., 2000), engineering (Zakariya et al., 2023), life sciences (Adkins et al., 2023; Leiva-Brondo et al., 2020), physiology (Thompson, 2024), and health sciences (LoGiudice et al., 2023). Additionally, the R-SPQ-2F provided a non-overlapping set of items with the M-ASSIST. It may be that the items on the R-SPQ-2F better reflect deep-understanding processes (deep approach to learning) that presumably lead to more abstract and encompassing cognitive representations and similarly better reflect surface learning processes (surface approach to learning) that support reliance on reproduction of the instructional examples (exemplar-based cognitive representations).

Again, in the present study, if Hypothesis 1 is supported, we should find significant associations between students' self-reported approaches to learning and their concept-building tendencies (abstractor, exemplar). Alternatively, if Hypothesis 2 has merit, then the same pattern reported in Studies 1a and 1b should be observed; i.e., no evidence of an association between students' self-reported approaches to learning and their objectively determined concept-building tendencies.

3.1. Method

3.1.1. Participants

Study 2 was conducted in a research laboratory at a mid-sized private research university in the midwestern United States across two semesters (fall 2023 and fall 2024). Participants elected to participate as one of several options to fulfill a research credit requirement for their psychology courses. Consent was provided by 157 participants, of which 57 (36%) were abstraction learners, 56 (36%) were exemplar learners, 37 (24%) were non-learners, 5 (3%) did not complete the concept-building task, and 2 (1%) did not complete the R-SPQ-2F. The project was approved by the university's institutional review board (IRB ID no. 202308131).

3.1.2. The Revised Study Process Questionnaire (R-SPQ-2F)

The R-SPQ-2F is a 20-item questionnaire designed for use by educators to evaluate the learning approaches of their students (Biggs et al., 2001; see Supplementary Materials for full set of survey items). The scale has deep approach and surface approach scales, each with 10-items. Response options range from 'never or only rarely true of me' to 'always or almost always true of me' on a five-point Likert scale. The deep approach and surface approach scales were scored by creating a sum value of responses for each scale. The possible range of values for each scale is from 10 to 50. Biggs et al. (2001) examined the extent to which the items fit a deep approach and surface approach factor structure and yielded model fit indices of CFI = 0.99 and SRMR = 0.02 for their data, suggesting a good fit to the two-factor structure. Standardized factor loadings ranged from 0.31 to 0.63 for the deep items, and 0.38 to 0.67 for the surface items, with all p's < 0.05.

3.1.3. Procedure

The overall procedure and concept-building task were identical to that used in Study 1b; however, the order and number of surveys completed by participants varied by semester. In fall 2023, participants completed the R-SPQ-2F and a few short demographic questions; whereas in fall 2024, the R-SPQ-2F was completed last out of five surveys, the fourth being the M-ASSIST (the other surveys focused on motivational and mindset factors for purposes unrelated to this study). Participants were allowed up to 90 min to complete the concept-building task and surveys as part of one session.

3.1.4. Analyses

The analysis plan was similar to that used in Study 1 with correlation coefficients and a mixed-factor ANOVA utilized to compare concept-building classification with learning approaches. The ANOVA was a 2(Concept-Building Approach: Abstractor vs. Exemplar) X 2(R-SPQ-2F: Deep Approach vs. Surface Approach). Power analyses were again conducted using G*Power 3.1.9.7 (Faul et al., 2007, 2009) to determine the power to detect medium and large effects. At a significance level of $p = 0.05$ with $N = 113$ (all abstraction and exemplar learners), the sample provided 90% power to detect a medium correlation and provided 91% power to detect a medium effect size when testing the interaction of concept-building and learning approach. There was 99% power to detect a large correlation and 99% power to detect a large effect size.

3.2. Results and Discussion

Table 3 displays correlation coefficients for concept-building approach (as point biserial correlations), extrapolation MAE (continuous index of concept-building tendency), R-SPQ-2F deep approach, and R-SPQ-2F surface approach scores. Concept-building approach was not related to the deep approach scale of the R-SPQ-2F or with the surface approach scale

of the R-SPQ-2F. Extrapolation MAE was also not related to the deep or surface scales. The deep and surface approach scales of the R-SPQ-2F were significantly negatively correlated with one another.

Table 3. Correlations for concept-building approach and R-SPQ-2F deep approach and surface approach scales.

Variable	1	2	3	4
1. Concept-Building Approach [a]	-			
2. Extrapolation MAE	0.88 *	-		
3. R-SPQ-2F Deep Approach	−0.11	−0.06	-	
4. R-SPQ-2F Surface Approach	−0.003	0.003	−0.28 *	-

Note. [a] 0 = abstraction learners, 1 = exemplar learners. * $p < 0.05$. $n = 113$ for all correlations.

Figure 5 displays the average scores for the deep approach and surface approach scales of the R-SPQ-2F for abstraction and exemplar learners. Significantly more deep approach items were endorsed by all learners compared to surface approach items, $F(1, 111) = 7.34$, $p = 0.008$, $\eta^2_p = 0.062$. Exemplar and abstraction learners did not significantly differ in their endorsement of items on the R-SPQ-2F, $F(1, 111) = 0.86$, $p = 0.356$, $\eta^2_p = 0.008$. Most importantly, concept-building tendency did not significantly interact with R-SPQ-2F learning approach, $F(1, 111) = 0.43$, $p = 0.514$, $\eta^2_p = 0.004$.

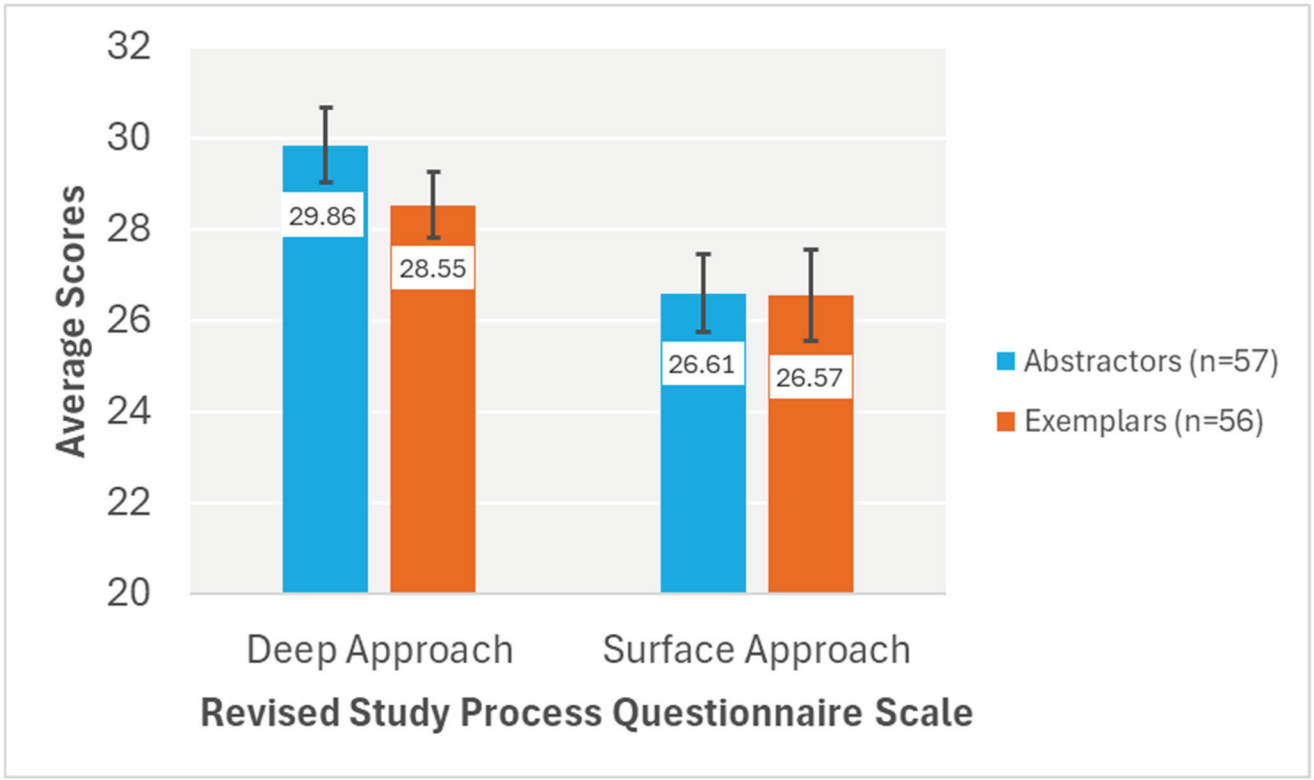


Figure 5. R-SPQ-2F deep approach and surface approach average scores for abstractor and exemplar learners. ***Note.*** Error bars represent standard error of the mean.

3.3. Exploratory Analysis

Following Thompson's (2024) comparison of the ASSIST and the R-SPQ-2F, as an exploratory analysis, we examined the relationship between the deep and surface approach items from the M-ASSIST and R-SPQ-2F using the fall 2024 subsample of participants, who completed both measures. Unlike the previous analyses, non-learners from the concept-building task were included in these correlations to achieve greater power, if they

fully completed both the M-ASSIST and the R-SPQ-2F. Table 4 displays the correlation coefficients between the deep and surface approach scales from both the M-ASSIST and the R-SPQ-2F. The deep approach scales of the M-ASSIST and R-SPQ-2F were significantly highly correlated, and the surface approach scales of both instruments showed a significant moderate-to-strong correlation. There was also a small-to-moderate significant negative correlation between the surface scale of the M-ASSIST and the deep approach scale of the R-SPQ-2F, as well as a significant small-to-moderate negative correlation between the deep scale of the M-ASSIST with the surface scale of the R-SPQ-2F. The direction and strength of these relationships largely mirror that found by Thompson (2024), with the exception that Thompson did not find a significant negative relationship between the surface scale of the ASSIST with the deep scale of the R-SPQ-2F. Overall though, our analysis extends Thompson's findings by using the more succinct M-ASSIST instead of the ASSIST and providing similar results with a different sample of participants.

Table 4. Correlations for M-ASSIST and R-SPQ-2F deep approach and surface approach scales.

Variable	1	2	3	4
1. M-ASSIST Deep Approach	-			
2. M-ASSIST Surface Approach	−0.083	-		
3. R-SPQ-2F Deep Approach	0.599 *	−0.294 *	-	
4. R-SPQ-2F Surface Approach	−0.378 *	0.400 *	−0.253 *	-

Note. * $p < 0.05$. $n = 77$ for all correlations; these include non-learners.

Study 2 expanded on Study 1 by employing the widely used R-SPQ-2F as a measure of student approaches to learning and replicated the previous findings examining the relation of concept-building tendency with self-reported approaches to learning (indexed by the M-ASSIST) in a laboratory setting. There was no significant relationship between the deep approach score from R-SPQ-2F and the abstraction concept-building tendency or between the surface approach score from R-SPQ-2F and the exemplar concept-building tendency. These findings again lend support for Hypothesis 2, that the concept-building tendency of a learner is not related to the learner's self-reported approaches to learning. To provide further evidence for Hypothesis 2, the Bayesian alternative to null hypothesis significance testing was again conducted as in Study 1. For the interaction in Study 2, the posterior probability for the null hypothesis being true was 0.895, providing "positive" evidence for the null hypothesis model (Raftery, 1995).

4. General Discussion

In this paper, we compared two frameworks used to measure students' learning approaches: the concept-building framework and the student approach to learning (SAL) framework. We wanted to explore potential overlap between these frameworks in revealing how students learn. To investigate this issue, we conducted two related studies in which college students completed both SAL self-report surveys and a function-learning task designed to determine college students' concept-building tendencies (abstraction vs. exemplar learning tendencies). For the SAL surveys, we used two common instruments in STEM education: M-ASSIST, an adaptation of ASSIST for chemistry, and R-SPQ-2F, which has been broadly used in many disciplines. These studies involved different student populations and were conducted in both a chemistry classroom and a research laboratory setting.

As developed in the introduction, it seemed plausible that these frameworks could be converging on a similar idea in individual differences among learners in the fundamental ways in which they approach learning challenges. However, across both studies, we found a consistent pattern: There was no detectable association between the deep versus surface

learning approaches of the SAL framework (as measured by two SAL inventories) and the abstraction versus exemplar approaches from the concept-building framework. Specifically, in both studies (Studies 1a and 1b combined and Study 2), we found no significant correlation between a student's concept-building approach and their SAL self-reports, whether using M-ASSIST or R-SPQ-2F. This lack of association remained consistent across both classroom and laboratory settings, despite the studies having sufficient power to detect medium size correlations. These findings suggest that the SAL framework instruments do not differentiate between exemplar and abstraction learners.

To further explore potential connections, we examined whether exemplar learners scored higher on surface approach items and abstraction learners scored higher on deep approach items. If the two frameworks measured a similar tendency, this pattern would be expected. However, across both SAL surveys, all students endorsed significantly more deep approach items than surface approach items, regardless of their concept-building approach. Thus, we found no empirical evidence linking these two theoretical frameworks. It could be that the lack of empirical overlap between the two frameworks stemmed from procedural, methodological, or participant-related factors affecting the behavior of the SAL inventories in our studies. Thompson (2024) found that in an undergraduate anatomy course, the R-SPQ-2F and ASSIST surveys aligned, with deep approach scores and surface approach scores having significant positive correlations across both instruments. This finding suggests that these two surveys are measuring similar constructs. If the M-ASSIST and R-SPQ-2F are functioning as expected in our studies, similar correlations should emerge. Our results confirmed this expectation. There were significant positive correlations between the M-ASSIST and R-SPQ-2F deep approach scores, as well as between their surface approach scores. Furthermore, we found significant negative correlations between deep and surface approaches across the two surveys. Hence, these findings support stability in our assessment of the deep and surface approaches, aligning with previous research. In the next section, we consider several alternative reasons why there was no support for an empirical overlap between the measurement indices that are commonly favored by these two frameworks.

4.1. No Empirical Overlap: Why?

One possible reason for no empirical overlap between the two frameworks is that the self-report inventories inspired by the SAL framework (Entwistle & McCune, 2004) may not provide a valid indicator of whether a particular student's learning reflects a push toward understanding the content (deep approach) or reflects a push toward reproducing that content (surface approach). Students' responses to some individual items could be inaccurate because they simply make careless responses or misinterpret the items. For instance, Mogashana et al. (2012) collected interviews from students who were instructed to elaborate on their responses to items on the Approaches to Learning and Studying Inventory (Entwistle & Ramsden, 1983) that target deep and surface approaches to learning. In some cases, in the interviews, students disagreed with their response(s), claiming that it was a mistake. In other cases, certain items confused students primarily because the entire statement was not well understood (e.g., "I just don't get this one. It is a complicated statement", p. 788), or because the students struggled in how to interpret a key word in the item (e.g., "trouble"). Additionally, interviews with undergraduates in an Anatomy class about their responses to the R-SPQ-2F survey found misalignments between their quantitative results and their qualitative data for at least 12 of the 20 items (Johnson et al., 2021). Students interpreted words in the items differently, and there were several compound items further complicating what exactly students were responding to.

In a related vein, these self-report inventories rely on the assumption that students' metacognitions regarding their learning are accurate. Cognitive and education stud-

ies have shown; however, that learners' metacognitions can be substantially inaccurate (Dunlosky & Metcalfe, 2008). With regard to the student approach to learning inventories, one challenge is that the student must be able to accurately report on how they learn. That is, students may believe that they are implementing a described activity, but in fact, they are not. These potential problems with the inventory content itself, along with possible metacognitive inaccuracies in how respondents report their approach to learning, may reduce the validity of the commonly used inventories for assessing student approaches to learning. As a result, this could weaken any empirical relationship between the SAL framework inventory assessments (e.g., the R-SPQ-2F) and the objective learning assessments used in the concept-building framework.

A second possible reason is that students' responses to self-report inventories could in part reflect their intended approach, but not what they are able to achieve. Findings in classroom contexts (e.g., Frey et al., 2020; Wilujeng & Alvarez, 2025) consistently demonstrate that learners are not especially good at judging their level of understanding. Algebra students can express high confidence in incorrect answers, unaware that their understanding of the principles underlying the problem is flawed (e.g., Wilujeng & Alvarez, 2025). Similarly, Frey et al. (2020), using think-aloud protocols, confirmed that students' incorrect answers on chemistry problems could be driven by incomplete understanding of the underlying concepts. Nevertheless, students were generally overconfident in the accuracy of these answers, suggesting that students were unaware that they had incomplete understanding of the requisite chemistry concepts. The point here is that some students may adopt a deep approach to learning—and thus endorse items on a SAL inventory that convey they are attempting to fit and relate ideas together, to understand the meaning of what they have to learn, to self -test until they have a complete understanding, and so on. Yet, because of poor meta-comprehension or because of having less skill at building organized cognitive structures (less-skilled structure builders (McDaniel et al., 2022)), these students do not achieve a deep understanding. Consequently, the association between responses on the SAL inventories and the objective learning patterns on the learning task used for identifying an individual's concept-building tendency may not be as robust as underlying theory might suggest it should be. This may especially be the case for the deep-approach scale, where for a number of reasons just reviewed, the scale may not discriminate among the students who actually do achieve deep understanding versus those who attempt to understand the concept but ultimately do not.

A third potential reason for our finding of no empirical relation across the SAL and concept-building frameworks hinges on some SAL theorists' emphasis that student approaches to learning are determined contextually (e.g., are dependent on specific course demands or content) (Sinapuelas & Stacy, 2015; Case & Marshall, 2004), which are revealed when students respond to the inventories within a particular course context. Because the concept-building task is neutral to course content and not completed with reference to particular course demands or content, concept-building approaches revealed by this task may not intersect with student approaches to learning revealed through SAL inventories administered within a course context (e.g., Study 1a). When SAL inventories are not administered within a course context (e.g., Study 1b, Study 2), it may be that they are not especially valid because they are not sensitive to the context dependencies that some SAL theorists believe are essential to understanding student approaches to learning.

In summary, our study found a disconnect between two well-known frameworks for understanding how students learn: the concept-building framework and the SAL framework. Despite theoretical overlap suggesting that deep learning approaches might align with abstraction-based concept-building and surface learning approaches might align with exemplar-based concept building, our empirical findings across two studies and multiple contexts did not support this assumption. The lack of correlation may reflect limitations

inherent to self-report inventories, including inaccurate metacognitive judgments and item misinterpretation, as well as the potential for a misalignment between students' intended and actual learning outcomes. Furthermore, contextual factors emphasized by SAL theorists may limit the utility of SAL inventories when used outside of specific course settings. These results suggest that while both frameworks offer valuable perspectives, they may be tapping into fundamentally different dimensions of student learning.

4.2. Study Limitations and Future Directions

The present study has some limitations worth acknowledging and also sets the stage for potential future directions into deepening our understanding of the relationship between these two frameworks, as well as how students learn. First, for classroom application, the concept-building task (function-learning) has some practical limitations. The task is relatively demanding, requiring about 45 min to complete. Moreover, to determine a student's concept-building approach, a high level of learning in the training phase must be achieved (otherwise transfer patterns are indeterminate of individual differences in abstraction vs. exemplar learning). Unfortunately, 23% or more students typically do not meet this criterion in classroom studies (21–24% in the present studies conducted). It appears that part of this "failure" rate hinges on these students exerting less effort than those that meet the learning criterion (McDaniel et al., 2022). Incentives to increase motivation to perform well in the training phase might alleviate this problem. Alternatively, there is an existing category learning task that distinguishes between abstraction and exemplar learners but is less demanding and exhibits a much lower "failure" rate (J. L. Little & McDaniel, 2015). Future work could explore the extent to which this task provides an index of individual differences in concept-building approaches that is a valid indicator of classroom performance. Additionally, future work could explore having learners write down their problem-solving steps during the training phase. For example, we might see exemplar learners writing down input-output pairs for the training examples, whereas abstraction learners may attempt to graph the underlying function. Incorporating this aspect into the task could potentially reduce the number of non-learners or, at the very least, provide deeper insight into how different learners engage with the task.

A second limitation is that we did not examine the accuracy of students' metacognitions regarding their responses to the approaches to learning inventories. Consequently, possible explanations that hinge on metacognition for why there was no empirical association between the SAL and concept-building frameworks remain speculative. Future work could examine the role of learners' metacognition in influencing the accuracy of their self-reported approaches to learning. Following the issues raised in the preceding section related to students' generally poor ability to estimate their own learning and judge their level of understanding, future studies might directly compare students' SAL inventory responses with these students' metacognitive accuracy. Two possibilities could be to (a) collect students' inventory responses and sometime later conduct think-alouds for these students as they read a scientific text to evaluate the degree to which they are indeed activating surface versus deep learning approaches; or (b) administer comprehension questions to evaluate the degree to which deep understanding was achieved (e.g., Gouravajhala & McDaniel, 2025; McNamara et al., 1996) as a function of inventory response patterns.

A third limitation is that we did not assess course performance for our participants, to try to gauge the degree to which there are linkages between course performances (e.g., exam scores) and individual differences in learning approaches or in concept-building approaches. Future work could fruitfully examine the relationship between course achievement (exam questions or exam scores) and learners' surface and deep approach tendencies on the one hand and their concept-building classification on the other hand. Existing research generated from each framework separately has done so, with somewhat divergent results. A deep

learning approach should theoretically correlate positively with performance on transfer (or higher-order) questions. Yet, Thompson (2024), in an undergraduate human anatomy course, found no significant correlation between ASSIST or R-SPQ-2F learning approach scores (deep or surface) and exam scores, even when distinguishing between lower-order and higher-order questions. Similarly, LoGiudice et al. (2023), in two undergraduate health sciences courses, found no significant correlation between the R-SPQ-2F deep approach scores and student performance on transfer questions. By contrast, as reviewed in the Introduction, individual differences in abstraction learning tendencies versus exemplar learning tendencies are consistently associated with course achievement (Frey et al., 2017), particularly when course exams include transfer questions (McDaniel et al., 2018, 2022). However, the courses and course assessments across studies adopting the SAL inventories relative to those adopting the concept-building approach learning task have not been all that similar. Future studies that examine how both frameworks fare in predicting exam performances for the same students in the same courses could provide a substantial advance (with exam questions that target retention and questions that target transfer of course material).

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/bs15081055/s1.

Author Contributions: Conceptualization, M.A.M., R.F.F. and H.K.B.; methodology, M.A.M., C.M.W. and H.K.B.; formal analysis, M.A.M. and C.M.W.; data curation, C.M.W. and H.K.B.; writing—original draft preparation, M.A.M., C.M.W. and R.F.F.; writing—review and editing, M.A.M., C.M.W. and R.F.F.; project administration, M.A.M.; funding acquisition, M.A.M. All authors have read and agreed to the published version of the manuscript.

Funding: Study 1a was supported by a grant to J. Mestre and M.A.M. from the National Science Foundation Division of Undergraduate Education, grant no. DUE-1630128, to the University of Illinois, Champaign–Urbana.

Institutional Review Board Statement: The studies were conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Boards of St. Louis College of Pharmacy (IRB ID no. 2017-39, 16 October 2017) and Washington University in St. Louis (IRB ID no. 201710100, 30 October 2017) and (IRB ID no. 202308131, 28 August 2024).

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: The data presented in this study are available upon request from the authors.

Acknowledgments: We would like to acknowledge the technical and programming support provided by Michael J. Cahill in setting up the function learning task used in the reported studies.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Notes

[1] Many commonly used SAL surveys in STEM are based on the Approaches and Study Skills Inventory for Students (ASSIST) inventory (Tait et al., 1997). The ASSIST survey has 52 items that are related to approaches to studying and learning. Deep learning items ask about seeking meaning in the task, relating ideas, understanding how ideas fit together, interest in the ideas, and monitoring effectiveness; surface learning items focus on lack of purpose, unrelated memorizing, and fear of failure. This survey has been used in STEM courses such as engineering (Rowe, 2002), nursing (Trigwell & Prosser, 1991), and chemistry (Brown et al., 2015). The survey also features items comprising a strategic component, relating to study organization, time management, motivation, and assessment demands.

[2] The observant reader may notice that there is a slight positive linear bias for exemplar learners seen in the left of the top chart and the right of the bottom chart of Figure 2, consistent with a ubiquitous bias in human function learning (see Busemeyer et al., 1997).

References

Adkins, S., Brasfield, K., Tran, T., Morris, J. J., & Raut, S. A. (2023). Which course resources and student approaches to learning are related to higher grades in introductory biology? *Journal of Research in Science, Mathematics, and Technology Education*, *6*(3), 175–196. [CrossRef]

Atieh, E. L., York, D. M., & Muñiz, M. N. (2020). Beneath the surface: An investigation of general chemistry students' study skills to predict course outcomes. *Journal of Chemical Education*, *98*(2), 281–292. [CrossRef]

Biggs, J. B. (1987a). *Student approaches to learning and studying*. Australian Council for Educational Research.

Biggs, J. B. (1987b). *The study process questionnaire (SPQ): Manual*. Australian Council for Educational Research.

Biggs, J. B. (1993). What do inventories of students' learning processes really measure? A theoretical review and clarification. *British Journal of Educational Psychology*, *63*(1), 3–19. [CrossRef]

Biggs, J. B., Kember, D., & Leung, D. Y. (2001). The revised two-factor study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, *71*(1), 133–149. [CrossRef]

Bourne, L. E. (1974). An inference model for conceptual rule learning 1. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium* (pp. 231–256). Routledge.

Bourne, L. E., Jr., Raymond, W. D., & Healy, A. F. (2010). Strategy selection and use during classification skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 500. [CrossRef]

Brown, S., White, S., Wakeling, L., & Naiker, M. (2015). Approaches and study skills inventory for students (ASSIST) in an introductory course in chemistry. *Journal of University Teaching and Learning Practice*, *12*(3), 1–14. [CrossRef]

Bunce, D. M., Komperda, R., Schroeder, M. J., Dillner, D. K., Lin, S., Teichert, M. A., & Hartman, J. R. (2017). Differential use of study approaches by students of different achievement levels. *Journal of Chemical Education*, *94*(10), 1415–1424. [CrossRef]

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts, & D. R. Shanks (Eds.), *Knowledge, concepts and categories* (pp. 408–437). The MIT Press.

Case, J., & Marshall, D. (2004). Between deep and surface: Procedural approaches to learning in engineering education contexts. *Studies in Higher Education*, *29*(5), 605–615. [CrossRef]

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968–986. [CrossRef]

dos Santos Rodrigues, M. N., & Gomes, C. M. A. (2020). Testing the hypothesis that the deep approach generates better academic performance. *International Journal of Development Research*, *10*(12), 42925–42935.

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Sage Publications.

Entwistle, N., Hanley, M., & Hounsell, D. (1979). Identifying distinctive approaches to studying. *Higher Education*, *8*, 365–380. [CrossRef]

Entwistle, N., & McCune, V. (2004). The conceptual bases of study strategy inventories. *Educational Psychology Review*, *16*(4), 325–345. [CrossRef]

Entwistle, N., & Ramsden, P. (1983). *Understanding student learning*. Croom Helm.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. [CrossRef]

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. [CrossRef]

Frey, R. F., Cahill, M. J., & McDaniel, M. A. (2017). Students' concept-building approaches: A novel predictor of success in chemistry courses. *Journal of Chemical Education*, *94*(9), 1185–1194. [CrossRef]

Frey, R. F., McDaniel, M. A., Bunce, D. M., Cahill, M. J., & Perry, M. D. (2020). Using students' concept-building tendencies to better characterize average-performing student learning and problem-solving approaches in general chemistry. *CBE—Life Sciences Education*, *19*(3), ar42. [CrossRef]

Goldwater, M. B., Don, H. J., Krusche, M. J. F., & Livesey, E. J. (2018). Relational discovery in category learning. *Journal of Experimental Psychology: General*, *147*(1), 1–35. [CrossRef]

Google, A. N., Gardner, G., & Grinath, A. S. (2023). Undergraduate students' approaches to learning biology: A systematic review of the literature. *Studies in Science Education*, *59*(1), 25–66. [CrossRef]

Gouravajhala, R., & McDaniel, M. A. (2025). Learning from educational texts: Is reduction in global text cohesion a desirable difficulty as structure building skill increases? *Learning and Individual Differences*, *121*, 102720. [CrossRef]

Gouravajhala, R., Wahlheim, C. N., & McDaniel, M. A. (2020). Individual and age differences in block-by-block dynamics of category learning strategies. *Quarterly Journal of Experimental Psychology*, *73*(4), 578–593. [CrossRef]

Herzog, S. A., & Goldwater, M. B. (in press). Evidence for rule versus exemplar learning strategies as stable individual differences independent from working memory. *Memory & Cognition*.

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). Pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology: General*, *143*(6), 2242. [CrossRef]

Johnson, S. N., Gallagher, E. D., & Vagnozzi, A. M. (2021). Validity concerns with the revised study process questionnaire (R-SPQ-2F) in undergraduate anatomy & physiology students. *PLoS ONE*, *16*(4), e0250600.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Juslin, P., Olsson, H., & Olsson, A. C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*(1), 133. [CrossRef]

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163. [CrossRef] [PubMed]

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. [CrossRef]

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 811. [CrossRef]

Lahdenperä, J., Postareff, L., & Rämö, J. (2019). Supporting quality of learning in university mathematics: A comparison of two instructional designs. *International Journal of Research in Undergraduate Mathematics Education*, *5*, 75–96. [CrossRef]

Lee, S. W. Y., Liang, J. C., & Tsai, C. C. (2016). Do sophisticated epistemic beliefs predict meaningful learning? Findings from a structural equation model of undergraduate biology learning. *International Journal of Science Education*, *38*(15), 2327–2345. [CrossRef]

Leiva-Brondo, M., Cebolla-Cornejo, J., Peiró, R., Andrés-Colás, N., Esteras, C., Ferriol, M., Merle, H., Díez, M. J., & Pérez-de-Castro, A. (2020). Study approaches of life science students using the revised Two-Factor Study Process Questionnaire (R-SPQ-2F). *Education Sciences*, *10*(7), 173. [CrossRef]

Li, W. T., Liang, J. C., & Tsai, C. C. (2013). Relational analysis of college chemistry-major students' conceptions of and approaches to learning chemistry. *Chemistry Education Research and Practice*, *14*(4), 555–565. [CrossRef]

Little, D. R., Nosofsky, R. M., & Denton, S. E. (2011). Response-time tests of logical-rule models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 1. [CrossRef] [PubMed]

Little, J. L., & McDaniel, M. A. (2015). Individual differences in category learning: Memorization versus rule abstraction. *Memory & Cognition*, *43*, 283–297.

LoGiudice, A. B., Norman, G. R., Manzoor, S., & Monteiro, S. (2023). Does a deep learning inventory predict knowledge transfer? Linking student perceptions to transfer outcomes. *Advances in Health Sciences Education*, *28*(1), 47–63. [CrossRef]

Maciejewski, W., & Merchant, S. (2016). Mathematical tasks, study approaches, and course grades in undergraduate mathematics: A year-by-year analysis. *International Journal of Mathematical Education in Science and Technology*, *47*(3), 373–387. [CrossRef]

Malinakova, H. C. (2022). Assessment of students' study approaches in the first semester of organic chemistry: Patterns of evolution diverge according to students' achievement level. *Journal of Chemical Education*, *99*(8), 2787–2797. [CrossRef]

Malinakova, H. C. (2024). Longitudinal study of students' study approaches in organic chemistry classes: Using structural equation modeling (SEM) to detect a response shift in data from the OCH-adjusted M-ASSIST instrument. *Journal of Chemical Education*, *102*(1), 15–26. [CrossRef]

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. *British Journal of Educational Psychology*, *46*(1), 4–11. [CrossRef]

Marton, F., & Säljö, R. (1997). Approaches to learning. In F. Marton, D. J. Hounsell, & N. J. Entwistle (Eds.), *The experience of learning* (2nd ed., pp. 39–58). Scottish Academic Press.

Masson, M. E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*, 679–690. [CrossRef]

May, W., Chung, E. K., Elliott, D., & Fisher, D. (2012). The relationship between medical students' learning approaches and performance on a summative high-stakes clinical performance examination. *Medical Teacher*, *34*(4), e236–e241. [CrossRef] [PubMed]

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words? *Journal of Educational Psychology*, *82*(4), 715. [CrossRef]

McDaniel, M. A., Cahill, M. J., Frey, R. F., Limeri, L. B., & Lemons, P. P. (2022). Learning introductory biology: Students' concept-building approaches predict transfer on biology exams. *CBE—Life Sciences Education*, *21*(4), ar65. [CrossRef] [PubMed]

McDaniel, M. A., Cahill, M. J., Frey, R. F., Rauch, M., Doele, J., Ruvolo, D., & Daschbach, M. M. (2018). Individual differences in learning exemplars versus abstracting rules: Associations with exam performance in college science. *Journal of Applied Research in Memory and Cognition*, *7*(2), 241–251. [CrossRef]

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*, *143*(2), 668. [CrossRef] [PubMed]

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and instruction*, *14*(1), 1–43. [CrossRef]

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207. [CrossRef]

Mogashana, D., Case, J. M., & Marshall, D. (2012). What do student learning inventories really measure? A critical analysis of students' responses to the approaches to learning and studying inventory. *Studies in Higher Education*, *37*(7), 783–792. [CrossRef]

Nelson Laird, T. F., Shoup, R., Kuh, G. D., & Schwarz, M. J. (2008). The effects of discipline on deep approaches to student learning and college outcomes. *Research in Higher Education*, *49*, 469–494. [CrossRef]

Nosofsky, R. M., & Kruschke, J. K. (1992). Investigations of an exemplar based connectionist model of category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation* (pp. 207–250). Academic Press.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the motivated strategies for learning questionnaire (MSLQ)*. The Regents of the University of Michigan.

Prosser, M., Trigwell, K., Hazel, E., & Waterhouse, F. (2000). Students' experiences of studying physics concepts: The effects of disintegrated perceptions and approaches. *European Journal of Psychology of Education*, *15*, 61–74. [CrossRef]

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. [CrossRef]

Rowe, J. W. (2002). First year engineering students' approaches to study. *International Journal of Electrical Engineering Education*, *39*(3), 201–209. [CrossRef]

Selcuk, G. S. (2010). The effects of problem-based learning on pre-service teachers' achievement, approaches and attitudes towards learning physics. *International Journal of the Physical Sciences*, *5*(6), 711–723.

Sinapuelas, M. L., & Stacy, A. M. (2015). The relationship between student success in introductory university chemistry and approaches to learning outside of the classroom. *Journal of Research in Science Teaching*, *52*(6), 790–815. [CrossRef]

Tait, H., Entwistle, N., & McCune, V. (1997). ASSIST: A reconceptualization of the approaches to studying inventory. In C. Rust (Ed.), *Improving student learning: Improving students as learners*. Oxford Brookes University Centre for Staff and Learning Development.

Thompson, A. R. (2024). A comparison of two learning approach inventories and their utility in predicting examination performance and study habits. *Advances in Physiology Education*, *48*(2), 164–170. [CrossRef]

Trigwell, K., & Prosser, M. (1991). Relating approaches to study and quality of learning outcomes at the course level. *British Journal of Educational Psychology*, *61*(3), 265–275. [CrossRef]

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wilujeng, H., & Alvarez, J. I. (2025). Students' misconceptions in algebraic concepts: A four-tier diagnostic test approach. *Jurnal Elemen*, *11*(1), 120–132. [CrossRef]

Zakariya, Y. F., Nilsen, H. K., Bjørkestøl, K., & Goodchild, S. (2023). Analysis of relationships between prior knowledge, approaches to learning, and mathematics performance among engineering students. *International Journal of Mathematical Education in Science and Technology*, *54*(6), 1015–1033. [CrossRef]

Zeegers, P. (2001). Approaches to learning in science: A longitudinal study. *British Journal of Educational Psychology*, *71*(1), 115–132. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Whether Interleaving or Blocking Is More Effective for Long-Term Learning Depends on One's Learning Strategy

Jeri L. Little [1,*] and Jexy A. Nepangue [1,2]

[1] Department of Psychology, California State University, East Bay, 25800 Carlos Bee Blvd, Hayward, CA 94542, USA; jnepangu@ucsc.edu
[2] Department of Psychology, University of California, 1156 High Street, Santa Cruz, CA 95064, USA
* Correspondence: jeri.little@csueastbay.edu

Abstract

Grouping information into categories enables us to learn, integrate, and apply new information. Presenting items from different categories sequentially (i.e., interleaving) is often more effective than presenting items from a single category sequentially (i.e., blocking), particularly when evaluating learning using memory-based tests. However, blocking can be more effective than interleaving for rule-based learning. Research has investigated optimal sequence (interleaving vs. blocking) for category learning when participants can either memorize or find rules, finding an interaction between learning strategy and sequence; that is, when memorizing, interleaving is more effective than blocking for classifying based on similarity, but when trying to find a rule, blocking is more effective than interleaving for classifying based on the rule. The goal of the present experiments was to replicate and extend this finding by examining transfer immediately after learning and then at a delay of about 48 h. The present results replicate the interaction between sequence and strategy, and show that the pattern persists over a delay. The results also suggest that some rule-based learning may be more resistant than memory-based learning to forgetting. These findings have educational implications for structuring learning as a function of strategy or learning goals.

Keywords: category learning; rule learning; blocking; interleaving

1. Introduction

Grouping information into categories enables us to learn, integrate, and apply new information. Students taking statistics classes need to learn when certain formulas and procedures are necessary. Students in geology courses learn to categorize rocks so that they can later classify specific examples they have not seen. Students often learn categories from examples, and when learning categories of information from examples, effective learning may depend on how those examples are sequenced. Much previous research has shown that presenting items from different categories sequentially (i.e., interleaving) is better for learning than presenting items from the same category sequentially (i.e., blocking; see reviews by Brunmair & Richter, 2019; Firth et al., 2021). However, this advantage of interleaving over blocking depends on the features of the materials and the task. For example, when there is a clear rule and/or items have low within-category similarity, blocking can outperform interleaving (e.g., Noh et al., 2016; Sorensen & Woltz, 2016). Additionally, participants can approach a single task with different strategies, and the optimal sequence can depend on the strategy a learner uses in a task (e.g., memorization

vs. rule finding)—at least on tests provided immediately after training (Little et al., 2025). The goal of the present study is to examine optimal sequence (interleaving vs. blocking) as a function of instructed strategy (to memorize or to find a rule) in long-term learning.

1.1. Optimal Sequence for Learning Depends on Strategy

Interleaving items from different categories is often better for learning than blocking those items by category (see reviews by Brunmair & Richter, 2019; Firth et al., 2021). This benefit has been shown in a variety of domains, including the learning of artists' styles (Kornell & Bjork, 2008), naturalistic stimuli (Wahlheim et al., 2011), math (Rohrer & Taylor, 2007), and simple motor skills (Wulf & Shea, 2002). In the case of visual information, the advantage of interleaving over blocking is often attributed to interleaving's affordance for discrimination—the discriminative contrast hypothesis (Birnbaum et al., 2013; Kang & Pashler, 2012). That is, when items in different categories have between-category similarity, seeing items from different categories in sequence allows a learner to pick out the critical features that differentiate categories (see also Carvalho & Goldstone, 2015, sequential attention hypothesis).

However, sometimes blocking is more effective than interleaving. For example, when items are defined by a rule that one needs to abstract but within-category similarity is low and/or between-category similarity is high, blocking items from that category is often advantageous because it allows learners to find the critical feature (or combination of features) that items in a category have in common (e.g., Carvalho & Goldstone, 2014b, 2015, 2017; Kurtz & Hovland, 1956). For example, Carvalho and Goldstone (2014b) used "blob" stimuli made up of curvilinear segments; a single curvilinear segment defined the category membership. They manipulated the similarity of items across categories to be high (i.e., blobs had largely the same shape within and between categories) or low (i.e., there were a variety of shapes within and between each category). They showed an interaction between sequence and similarity for performance on novel transfer items. Specifically, there was an advantage of interleaving over blocking for transfer classification when similarity was high, but an advantage of blocking over interleaving when similarity was low. In addition to interacting with the level of between- and within-category similarity, sequence has been shown to interact with other variables to influence learning. For example, Noh et al. (2016) found that sequence interacts with category structure; with a rule-based category structure, blocking outperformed interleaving, but with an information-integration-based category structure (i.e., that would rely more on exemplar-based than rule-based processes), interleaving outperformed blocking.

Related to the research by Noh et al. (2016), researchers have also found sequence to interact with the strategy that a learner uses to learn in a single task (Goldwater et al., 2018; Little et al., 2025). Specifically, in a series of experiments, Little et al. (2025) had participants study letter strings from four different categories in either a blocked or an interleaved order. The letter strings were individualized, but their category designation followed a rule pertaining to the last letter (e.g., items that ended with "t" were plants; those that ended with "k" were tools; those that ended with "b" were clothing; those that ended with "p" were animals). In the blocked condition, participants saw all of the items from a given category sequentially (e.g., plant—spaksvot, plant—coteset, plant—quitinoct). In the interleaved condition, participants saw items from a given category intermixed with items from the other categories (e.g., plant—spaksvot, animal—hazamap, tool—blasimark). On a transfer task, participants had to classify items that looked similar to the trained items but for which the last letter would designate a different category on the basis of the rule (e.g., spaksvok). The idea, based on the research by Little and McDaniel (2015a), was that memorizers would classify the items based on similarity (e.g., spaksvok is a plant), but rule

finders would classify the items based on the rule (e.g., spaksvok is a tool). Indeed, Little et al. found that participants differed in how they classified the items. Critically, when participants self-reported as memorizers, interleaving was more effective than blocking for later classification based on similarity, but when participants self-reported as rule finders, blocking was more effective than interleaving for later classification based on the rule. In Little et al.'s Experiment 2, they instructed participants to memorize or find a rule, and they observed the same pattern (participants were only given strategy instructions before training, not again before the test). That is, for those instructed to memorize during training, later similarity-based classification was facilitated by interleaving compared to blocking, but for those instructed to find a rule during training, later rule-based classification was facilitated by blocking compared to interleaving.

1.2. Category Learning over a Delay

Some research has examined the efficacy of interleaving versus blocking for long-term learning, examining performance on tests at delays of at least 24 h (Carvalho & Goldstone, 2014a, 2017; Dobson, 2011; Eglington & Kang, 2017; Linderholm et al., 2016; Rohrer et al., 2015; Zulkiply, 2013; Zulkiply & Burt, 2013). The general finding is that interleaving tends to outperform blocking, and while performance tends to be worse (overall) after a delay, the magnitudes of the effects shown immediately and at a delay tend to be comparable (i.e., no interaction between sequence and delay, see Firth et al., 2021). For example, Rohrer et al. (2015) examined interleaved versus blocked practice for math problems (graphs and slopes) and examined performance at one day and 30 days. They found an advantage of interleaving over blocking, and that effect was not significantly different at a 30-day delay. Similarly, Zulkiply (2013) examined interleaved versus blocked practice for texts presented aurally. They found an advantage of interleaving over blocking, and that effect was not significantly different at the one-week delay. Eglington and Kang (2017) examined interleaved versus blocked practice for categories of chemicals with a test delayed by two days (no immediate test). They found that interleaved practice outperformed blocked practice on a test at a two-day delay.

As mentioned in the earlier section, interleaving does not always outperform blocking on immediate tests, and this has implications for performance on delayed tests. Carvalho and Goldstone (2014a) examined both immediate and delayed transfer using the previously mentioned "blob" stimuli. Their Experiment 1a showed that although high-similarity items benefitted from interleaving over blocking for transfer performance on an immediate test, low-similarity items did not (the blocking advantage shown by Carvalho & Goldstone, 2014b, was not replicated). Nevertheless, the pattern of results was largely consistent for transfer items on the delayed test (24 h), again failing to reveal an interaction between sequence and delay (but see Linderholm et al., 2016, who did show evidence of an interaction between sequence and delay when combined with a testing manipulation).

Thus, most of the time, although performance is lower on a delayed test than on an immediate test, the relative effects of interleaving versus blocking shown on delayed tests are comparable to the effects shown on immediate tests. However, to our knowledge, none of the studies investigating category learning at a delay have used a rule-based category structure (and a task necessitating abstraction of that rule) that resulted in blocking outperforming interleaving on an immediate test. Thus, it is critical to know whether superior learning under blocked conditions would persist over time.

One can gain some insight into the retention of rule-based learning from the problem-solving literature. Abstracting a rule is like solving a problem. Evidence suggests that if a learner solves a problem, their memory for the solution is, to some extent, protected from forgetting (Dominowski & Buyer, 2000). Specifically, Dominowski and Buyer (2000)

compared those who solved problems to those who were given the solution. A week later, the "solvers" remembered the solution better than the participants who were given the solution (although see Weisberg & Alba, 1981, who showed forgetting over time with the nine-dot problem). First, Dominowski and Buyer's results suggest that rule-based performance may not be worse at a delay (which may be different from memory-based category learning). Second, if performance is not worse at a delay, it is also unlikely that the advantage of blocking over interleaving would change from an immediate to a delayed test, but this assumption warrants testing.

In summary, although performance in category learning tasks is often lower on a delayed test than on an immediate test, the relative effects of interleaving versus blocking shown on delayed tests tend to be comparable to those shown on immediate tests. However, to our knowledge, none of the studies investigating category learning at a delay have used a rule-based category structure and a task necessitating abstraction of that rule that resulted in blocking outperforming interleaving. Thus, it is critical to know whether superior learning under blocked conditions would persist over time.

1.3. The Present Research

The present research examines the interaction effect between sequence and instructed strategy on performance immediately and after a delay of about 48 h. The present Experiment 1 intends to replicate Little et al.'s (2025) Experiment 2 with a different sample and experimental context (mix of student and non-student participants, online vs. student participants, in-person). The present Experiment 2 extends the research by examining whether the effects persist over a delay, and how performance changes after re-training.

In the present experiments, participants studied letter strings from four different categories (from Little et al., 2025, experiments) in either an interleaved or blocked sequence. Prior to training, participants were either instructed to memorize or to find a rule. At the end of the first session, participants classified transfer items (without any further strategy instructions). In the present Experiment 2, during a second session about 48 h later, participants again classified the transfer items. Participants then completed a re-training phase. Finally, they classified those same transfer items for a third time.

We were most interested in performance on transfer items. First, following from research by Little et al. (2025), we expected that those told to memorize would tend to classify transfer items based on similarity to trained items, and those instructed to find a rule would tend to classify transfer items based on the rule. Second, for those instructed to memorize, interleaving would be more effective than blocking for classifying those transfer items based on similarity; for those instructed to find a rule, blocking would be more effective than interleaving for classifying based upon the rule. Finally, for those instructed to memorize, we expected the benefit of interleaving over blocking to persist when tested later, consistent with prior literature (Firth et al., 2021), but the predictions for rule-abstraction over time are less clear. Borrowing from the problem-solving literature, the rule—once found—may not be easily forgotten (Dominowski & Buyer, 2000). Another idea that would predict little forgetting is that figuring out a solution leads to deep encoding (e.g., Jacoby, 1978). In either case, we would expect to see little forgetting over time, and no interaction between sequence and delay for rule finders. However, although the general rule (i.e., last-letter-designated category membership) may be protected from forgetting, we wagered that specifics of the rule (e.g., "t" means plant, "b" means clothing) could be forgotten or mixed up, resulting in apparent forgetting on a delayed test. Thus, we added a re-training phase, followed by a third transfer test. If forgetting occurred, the responses during retraining would provide insight into the speed of relearning (i.e., given

a refresher; c.f. Carvalho & Goldstone, 2014a). We expected that the advantage of blocking over interleaving would remain after re-training.

2. Experiment 1

2.1. Method

2.1.1. Participants and Design

We recruited 201 participants (96 females, 97 males, 2 non-binary; average age = 32.5, *SD* = 11.84, range = 17 to 64) from psychology courses at California State University, East Bay (n = 62) and MTurk (www.mturk.com, n = 139). Students received 1 point towards their Research Participation Requirement at California State University, East Bay, for every 50 min (or any part thereof) they participated in. MTurk participants received USD 3 for their participation.

Little et al.'s (2025) Experiment 2, which most closely resembles the present experiment, revealed a large interaction effect (n_p^2 = 0.10), with an effect size for those instructed to memorize of d = 0.65 and for those instructed to find a rule of d = 0.64. In order to have 80% power to observe an effect size of d = 0.65 (alpha = 0.05), we needed 39 participants per group (156 total; G*Power 3.1, Faul et al., 2007). We collected data from extra participants to allow for cases we would need to remove. To maintain data integrity, we included attention checks, a training test to assess whether participants were trying to learn, and multiple open-ended questions to help distinguish real participants from bots and to help identify duplicate responses.

This experiment utilized a 2 × 2 between-subjects design. The independent variables were instructed strategy (find rule vs. memorize) and learning sequence (interleaved vs. blocked). Performance for transfer items was assessed based on participants' classification according to the instructed strategy during the training phase (i.e., memorization strategies were assessed for similarity-based transfer; rule-abstraction strategies were assessed for rule-based transfer).

2.1.2. Materials

Sixteen letterstrings previously used by Little et al. (2025) were used in the present experiment. There were four categories, with four letter strings in each category. The last letter of each letter string determined its category. Plants ended in "t" (i.e., quitinoct, ollinat, spaksvot, coteset). Tools ended in "k" (i.e., lontsink, yallinuk, martwok, blasimark). Clothing ended in "b" (i.e., jollinub, rigelwib, thadtelb, zabarab). Animals ended in "p" (e.g., grasavep, hazamap, flavanop, peetikip).

Transfer items were created to look identical to trained items, except that the last letter was switched with that from another category. Specifically, "t" and "k" were always switched, and "p" and "b" were always switched to minimize the difference in how the words sounded if pronounced. For example, for the trained item "plant—quitinoct", the transfer item would be "quitinock". The idea is that the participants who received the rule instructions during training would classify words based on the last letter (e.g., quitinock would be identified as a tool). Those who received the memorization instructions would classify based on the beginning of a word (e.g., quitinock would be identified as a plant).

2.1.3. Procedure

Participants completed this experiment on their own devices. They were instructed to learn words from four categories: plants, animals, tools, and clothing. They were told that they would see each word six times to help facilitate their learning, and that attention check trials would occasionally occur to check their engagement. They were also told not to write anything down. Participants were randomly assigned to a strategy condition (find rule or

memorize). Participants in the find rule condition were given the following instructions pertaining to their strategy:

> "Rule-based strategies involve trying to come up with a reason for why items are in each category (e.g., trying to find a pattern in letters, sounds, etc. that shows up in all of the words in the category, breaking the code)."

Participants in the memorization condition were told the following pertaining to the strategy they should use:

> "Memorization strategies might involve repeating an item over and over to yourself or coming up with a trick for remembering an item (e.g., associating something you already know with those new words, making up a rhyme to remember all of the items in a category, coming up with acronyms using the first letters of the items in the category, etc.)."

Participants were asked to re-state the instructions and then provided with a summary of the instructions. After the instruction manipulation, participants were randomly assigned to the interleaved or blocked sequence condition. In the interleaved sequence, the 16 items from the four different categories were presented in a mixed format (e.g., plant—spaksvot, animal—hazamap, tool—blasimark). Participants saw the 16 items in six blocks, with the items presented in a random order for each block. Starting in the third block, participants were given attention check trials (four in each block, one for each item). For the attention check trials, participants had to provide the category of an item presented the item before the last by choosing among the four categories. After choosing their answer, they were told the correct category designation. In the blocked condition, training consisted of six blocks of the four items from a given category, then six blocks of the four items from another category, and so forth, until all four categories were presented. Attention checks began in the third block for each category. In both conditions, items were shown for 6 s each time they were presented. Attention checks were self-paced.

After the training session, participants completed the training test in which they were instructed to classify the items they had learned by selecting one of the four trained categories. The test was self-paced, and no feedback was provided. Next, participants completed the transfer test in which they were instructed to classify new items. They were not told whether they should classify the items based on similarity or the rule, but our assumption based on previous research was that they would tend to classify based on the instructions they were given for training (Little et al., 2025). The test was self-paced, and no feedback was given. After participants completed the transfer test, they were told that they could have implemented a variety of strategies in this task. Participants were then asked to self-report their strategy on a scale of 1–7, where 1 = *solely memorization* and 7 = *solely trying to find a rule.* Lastly, participants were asked to describe a reason for the grouping of words if any were found.

At the end of the session, participants' demographic information was collected. They were asked if they had taken part in a study with similar materials and if yes, to explain. They were also asked about distractions, computer problems, and whether they wrote any information down. Most importantly, participants were reassured that their answers would not affect their compensation.

2.2. Results and Discussion

The data for the experiments in this paper, including the reasons for participant removal, can be found at Open Science Framework (OSF) (https://osf.io/z8b69/?view_only=71fa1b74f58a490a84ccc6b73f8a82d1, data posted 28 March 2025).

We removed 34 participants (12 CSUEB students, 22 MTurkers) from the analysis because they reported writing things down, reported having done the task before, their answer to the instruction summary and/or strategy question appeared to be copied and pasted from the internet or did not answer the question, or it appeared that they had done the study more than once (due to similar responses and/or IP addresses). This left 167 participants, with 38 in the rule–interleaved condition, 45 in the rule–blocked condition, 41 in the memorize–interleaved condition, and 43 in the memorize–blocked condition.

2.2.1. Manipulation Check

At the end of the task, we asked participants to rate their strategy on a 1–7 Likert scale where 1 = *solely memorization* and 7 = *solely trying to find a rule*. To assess the instruction manipulation's effectiveness, we conducted an independent-samples *t*-test on self-reported strategies as a function of instructed strategy. We found that those told to abstract the rule provided numerically—but not significantly—higher ratings ($M = 4.6$, $SD = 2.3$) than those told to memorize ($M = 4.1$, $SD = 2.1$), $t(165) = 1.51$, $p = 0.133$. Another way to assess whether the instruction manipulation succeeded was to examine the relationship between instructions and the classification of transfer items. Controlling for sequence condition, instructions (coded 1 for memorize, 2 for find rule) predicted rule-based categorization, $r(164) = 0.24$, $p = 0.002$, and similarity-based categorization, $r(164) = -0.24$, $p = 0.002$.

2.2.2. Performance for Trained Items

We conducted a 2 (strategy: find rule vs. memorize) × 2 (sequence: interleaved vs. blocked) ANOVA on performance for trained items. Participants performed numerically—but not significantly—better on the trained items if they had received the interleaved schedule ($M = 0.85$, $SE = 0.03$) than the blocked schedule ($M = 0.79$, $SE = 0.02$), $F(1, 163) = 2.40$, $p = 0.124$, $\eta_p^2 = 0.01$. There was no main effect of instructed strategy (find rule, $M = 0.83$, $SE = 0.02$; memorize, $M = 0.81$, $SE = 0.02$), $F(1, 163) = 0.74$, $p = 0.390$, $\eta_p^2 = 0.01$, and there was no interaction, $F(1, 163) = 0.78$, $p = 0.380$, $\eta_p^2 = 0.01$.

2.2.3. Performance for Transfer Items

The most critical measurement for the present paper was performance on transfer items. Critically, correct performance on transfer items reflected participants' instructions for training. For those instructed to find a rule during training, correct classification was based on adherence to the rule. For those instructed to memorize during training, the correct classification of transfer items was based on similarity to the trained item (i.e., same letter stem). As demonstrated in Figure 1, there appears to be an interaction between strategy instructions and sequence such that for rule finders, blocking is more effective than interleaving, but for memorizers, interleaving is more effective than blocking.

We analyzed these transfer data using a 2 (strategy: find rule vs. memorize) × 2 (sequence: interleaved vs. blocked) ANOVA. Sequence (interleaved vs. blocked) did not significantly affect correct classification, $F(1, 163) = 0.43$, $p = 0.515$, $\eta_p^2 = 0.003$ [interleaved $M = 0.54$, $SE = 0.04$; blocked $M = 0.51$, $SE = 0.04$]. However, those instructed to memorize did significantly better at correctly classifying the transfer items based on similarity ($M = 0.65$, $SE = 0.04$) than did those instructed to find a rule classify items based on the rule ($M = 0.40$, $SE = 0.04$), $F(1, 163) = 18.68$, $p < 0.001$, $\eta_p^2 = 0.10$.

Importantly, we found a significant interaction between strategy and sequence, $F(1, 163) = 9.16$, $p = 0.003$, $\eta_p^2 = 0.05$. Using independent-samples *t* tests, we conducted planned paired comparisons to assess the effectiveness of interleaving versus blocking for those instructed to memorize and those instructed to find a rule. For those instructed to memorize, interleaving was significantly more effective ($M = 0.76$, $SD = 0.30$) than blocking

(M = 0.54, SD = 0.36) for classifying based on similarity, $t(82) = 2.99$, $p = 0.002$, $d = 0.65$. For those instructed to find a rule, blocking was numerically—albeit not significantly—more effective (M = 0.47, SD = 0.44) than interleaving (M = 0.33, SD = 0.41) for classifying based on the rule, $t(81) = 1.50$, $p = 0.138$, $d = 0.33$).[1]

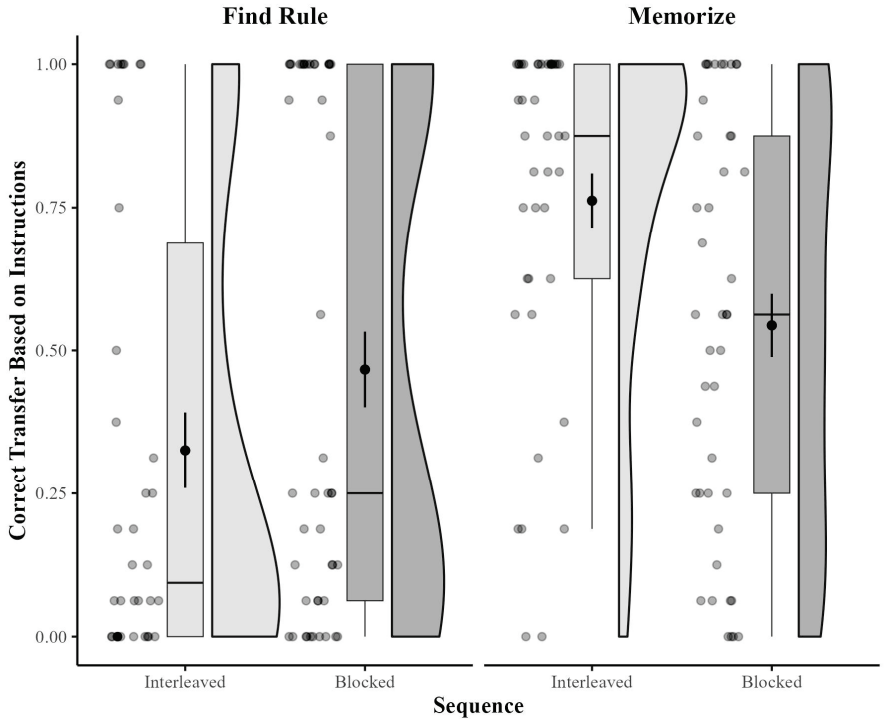


Figure 1. Correct categorization of transfer items as a function of instructed strategy and sequence in Experiment 1. Box plots display the interquartile range (IQR), with the whiskers extending to the minimum and maximum values within 1.5 times the IQR. The black circles within the box plots represent the mean for each cell, with the error bars representing +/−1 SEM. Violin plots depict the density of individual participants' accuracy scores, with wider sections indicating higher density. Individual participant scores are provided as a scatterplot to the left of the box plots.

It should be noted that across conditions, answers that were correct based on the rule accounted for 31% of the responses, and answers that were correct based on similarity accounted for 56% of responses. Answers that were errors (i.e., the two remaining possible responses) accounted for only 13% of the responses (chance would be 50%), suggesting that even when participants did not follow instructions, they seemed to classify based on either memory or the rule the majority of the time, which might reflect individual differences in predisposition to memorize or find rules that superseded the instructions, a point we will return to in the General Discussion.

Although MTurk participants generally outperformed the students, $F(1, 167) = 5.98$, $p = 0.016$, $\eta_p^2 = 0.04$, participant type did not interact with sequence ($p = 0.515$), strategy ($p = 0.324$), or the interaction of sequence and strategy ($p = 0.869$) to influence performance on transfer items.

2.2.4. Discussion

The results of Experiment 1 mostly replicated those reported by Little et al. (2025). Most critically, there was an interaction between sequence and strategy such that for those instructed to memorize, interleaving was more effective than blocking, but for those instructed to find a rule, blocking was numerically more effective than interleaving, although in the present study, it was a non-significant small-to-medium effect (d = 0.33). The finding that interleaving was more effective than blocking for classification based on

similarity is consistent with previous research (e.g., Kornell & Bjork, 2008). The finding that blocking is numerically more effective than interleaving for classification based on an articulable rule is also consistent with past research (e.g., Noh et al., 2016). However, the finding that both patterns could be present within the same materials—simply as a function of instructions—is more novel.

The goal of Experiment 2 was to examine whether this interaction would replicate and persist over a delay with a larger sample that would reveal a significant benefit of blocking over interleaving for those instructed to find a rule. To our knowledge, this would be the first study to examine the persistence of rule-based classification over time.

Consistent with prior literature (Firth et al., 2021), we would expect that for those instructed to memorize, interleaving would be more effective than blocking, and this effect would be comparable on a delayed test. However, the predictions for those instructed to find a rule were less clear because no research has investigated the persistence of rule-based classification over time, at least not in a task in which the participant needed to find the rule. We hypothesized that the rule may not be easily forgotten (c.f., problem-solving, Dominowski & Buyer, 2000), and even if the specifics of the rule were (e.g., which letter indicated each category), relearning might be faster for participants who had initially had blocked instruction than for those who had initially had interleaved instruction. For this reason, in Experiment 2, we included a transfer test at the beginning of the second session and then again after a re-training session.

3. Experiment 2

3.1. Method

3.1.1. Participants

Participants were recruited from MTurk (www.mturk.com). In the first session, 400 individuals participated (148 female, 249 male, 1 non-binary, 2 failed to report), aged 20–67 (M = 36.4, SD = 9.9). They were paid USD 3 for their participation in the first session.

The 319 participants who classified more than 50% of the trained items correctly at the end of Session 1 (regardless of other problems with their responses) were recruited to return two days later (81 were not asked to return). We recruited participants based on training performance because we believed participants learning fewer than 50% were likely bots or participants who were mindlessly guessing, and we initially believed that this would be a sufficient quality check given that we needed to deploy the task for the second session within 24 h of the end of data collection for the first session. We later removed more participants based on open-ended responses and other criteria, as discussed in the results section. Participants were incentivized to return for the second session with a notification when the study's second session was ready (in a note attached to a USD 0.50 bonus), and the 293 participants who returned were paid an additional USD 3 for their participation.

From the whole sample of 400 participants (including those who were not invited to return or who were invited but did not return), we removed 108 from the analyses because they reported writing things down, reported having done the task before, their answer to the strategy questions and/or question about instructions appeared to be copied and pasted from the internet or did not answer the question, or it appeared that they had done the study more than once (due to similar responses and/or IP addresses). This left 61 participants in the rule–interleaved condition (n = 50 who returned for Session 2), 72 in the rule–blocked condition (n = 57 who returned), 79 in the memorize–interleaved condition (n = 63 who returned), and 80 in the memorize–blocked condition (n = 58 who returned).

Power considerations. We aimed to reveal reliable differences in sequence for both those instructed to find a rule and those instructed to memorize. Little et al.'s (2025) Experiment 2 showed an effect size for those instructed to find a rule of d = 0.64 and for those

instructed to memorize of $d = 0.65$, but Experiment 1 in the present paper revealed effect sizes for those instructed to find a rule of $d = 0.33$ and for those instructed to memorize of $d = 0.65$. The lower average effect size was shown for rule finders ($d = 0.49$). To accommodate a smaller effect size, we modified our power calculations. Thus, to achieve 80% power (alpha = 0.05, two-tailed) to detect an effect of $d = 0.49$, we aimed for 67 participants per group. Unfortunately, we underestimated how many participants would need to be removed, so we ended up with less than 80% power for rule-finders for Session 2. Nevertheless, post-hoc power analyses revealed that given the sample that remained for Session 1, we had 80% power to detect an effect of $d = 0.49$ (alpha = 0.05, two-tailed) in the rule condition. Given the sample that remained for Session 2, we had 71% power to detect an effect size of $d = 0.49$ (alpha = 0.05, two-tailed) in the rule condition. We had 97% power to detect an effect of $d = 0.64$ in the memorize condition for Session 1 and 94% power to detect that effect for Session 2.

3.1.2. Design

This experiment utilized a 2 (strategy: find rule vs. memorize) × 2 (sequence: interleaved vs. blocked) × 3 (time: Session 1 vs. Session 2—beginning vs. Session 2—end) mixed-subjects design. As in Experiment 1, we manipulated instructed strategy (find rule vs. memorize) and learning sequence (interleaved vs. blocked) between subjects. Additionally, we assessed transfer performance at three points (within subjects): the end of Session 1, the beginning of Session 2 (about 48 h later), and the end of Session 2 (after another round of training). As in Experiment 1, participants were only given instructions for training, with the expectation that those instructions would influence how they classified items during the three transfer tests. Thus, as in Experiment 1, performance was assessed based on the classification of transfer items according to the instructed strategy (i.e., rule-finding strategies were assessed based on rule-based transfer; memorization strategies were assessed based on similarity-based transfer).

3.1.3. Materials and Procedure

Session 1. The materials were the same as those used in Experiment 1. Participants completed this experiment on their own devices. The procedure for Session 1 was the same as that used in Experiment 1 (training, test for training items, test for transfer items), except that participants did not rate their strategies at the end of Session 1 because we did not want to alert them to the other strategy. Instead, participants were given an open-ended question, which differed between the two instruction conditions. Participants in the rule condition were asked if they figured out the rule. Participants in the memorization condition were asked to explain what they did to memorize the items—or, if they did something different—to explain that strategy. We collected 99% of the responses within 24 h of deploying Session 1.

Session 2. Participants who answered more than 50% of training items correctly at the end of Session 1 were recruited to participate in Session 2. Session 2 was deployed 48 h after we deployed Session 1. Within an hour of deployment, we sent out a bonus of USD 0.50 with a message indicating that the second session of the study was ready to complete, and data collection occurred for up to 48 h. Ninety-eight percent of participants returned within 72 h. The average delay between the start of the first session and the start of the second session was 49 h (range 24 h to 85 h).

In the second session, participants were first tested on the 16 transfer items (self-paced, no feedback). Then, they completed re-training trials (four cycles, all in an interleaved sequence). Instead of presentation trials (i.e., used in Session 1), the re-training trials were test–feedback trials akin to the attention checks used in Session 1. Participants were

presented with the trained items and asked to choose the correct category. After choosing the answer, the correct category was provided.

After the re-training phase, participants completed the third transfer test with all 16 transfer items. After the transfer test, they provided a rating pertaining to their strategy (1 = *memorization*, 7 = *find a rule*). Specifically, they were asked, "On a scale of 1 to 7, where '1' means that you were relying solely upon memorization of the items and '7' means that you were solely trying to find a rule, give a rating for what you were doing when you were learning the items on the first day of this experiment?"

At the end of the session, participants' demographic information was collected. Then, they were asked if they had taken part in a study with similar materials and, if yes, to explain. As this was an online study, they were also asked about any distractions and computer problems they experienced while partaking in the study as well as whether they wrote anything down. Most importantly, they were reassured that their answers would not affect their compensation.

3.2. Results

3.2.1. Manipulation Check

At the end of Session 2, we asked participants to rate their strategy on a 1–7 Likert scale. Indeed, participants who were instructed to find a rule provided ratings that were more rule-like ($M = 5.0$, $SD = 2.1$) than did participants instructed to memorize ($M = 3.6$, $SD = 2.3$), $t(225) = 4.71$, $p < 0.001$, $d = 0.62$. Similar to our findings in Experiment 1, instructions also predicted how participants classified transfer items. Controlling for sequence condition, we found that instructed strategy (coded memorize = 1, find rule = 2) was associated with rule-based classification, $r(289) = 0.31$, $p < 0.001$, and similarity-based classification, $r(289) = -0.26$, $p < 0.001$, on the transfer task at the end of Session 1.

3.2.2. Session 1 Performance

Performance for trained items. We conducted a 2 (strategy: find rule vs. memorize) × 2 (sequence: interleaved vs. blocked) ANOVA on performance for trained items. Participants performed better on the training items if they had received the interleaved schedule ($M = 0.87$, $SE = 0.02$) compared to the blocked schedule ($M = 0.81$, $SE = 0.02$), $F(1, 288) = 4.90$, $p = 0.028$, $\eta_p{}^2 = 0.02$. There was no main effect of instructed strategy (memorize, $M = 0.82$, $SE = 0.02$; find rule, $M = 0.85$, $SE = 0.02$), $F(1, 288) = 1.33$, $p = 0.249$, $\eta_p{}^2 = 0.005$, and there was no interaction, $F(1, 288) = 0.91$, $p = 0.342$, $\eta_p{}^2 = 0.003$.

Performance for transfer items. Critically, we wanted to examine the potential interaction between strategy and sequence at the end of Session 1 (i.e., to replicate Experiment 1). As shown in Figure 2, we found an interaction such that when instructed to find a rule, participants tended to classify correctly based on the rule more often when they had received a blocked sequence than when they had received an interleaved sequence. However, when instructed to memorize, participants tended to classify correctly based on similarity more often when they had received an interleaved sequence than when they had received a blocked sequence. It also appeared that participants instructed to memorize were better at their task (i.e., classifying based on similarity) than rule-finders (i.e., classifying based on the rule).

To test these apparent results, we then conducted a 2 (strategy: find rule vs. memorize) × 2 (sequence: interleaved vs. blocked) ANOVA to assess the effects of strategy and sequence on correct transfer at the end of Session 1 based on the instructions. Indeed, we found a significant main effect of strategy such that participants instructed to memorize classified correctly based on similarity more ($M = 0.69$, $SE = 0.03$) than participants instructed to find a rule classified correctly based on the rule ($M = 0.42$, $SE = 0.03$),

$F(1, 288) = 38.89$, $p < 0.001$, $\eta_p^2 = 0.12$. There was also a main effect of sequence, such that participants classified items correctly more often when they had been blocked ($M = 0.61$, $SE = 0.03$) than when they had been interleaved ($M = 0.50$, $SE = 0.03$), $F(1, 289) = 6.44$, $p = 0.012$, $\eta_p^2 = 0.02$.

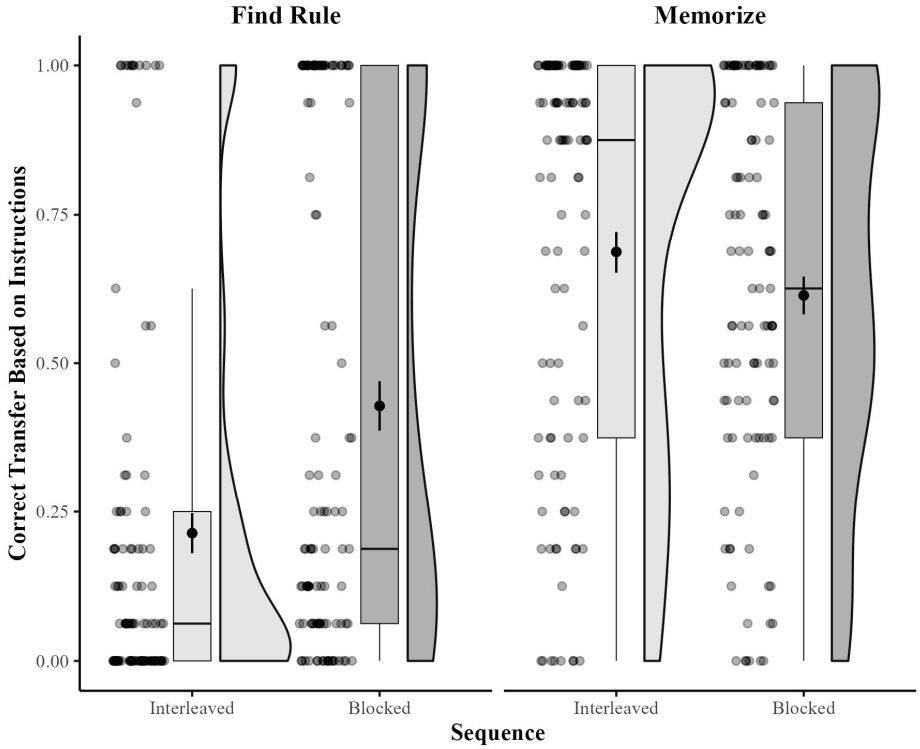


Figure 2. Correct categorization of transfer items as a function of instructed strategy and sequence at the end of Session 1 in Experiment 2. Box plots display the interquartile range (IQR), with the whiskers extending to the minimum and maximum values within 1.5 times the IQR. The black circles within the box plots represent the mean for each cell, with the error bars representing +/− 1 SEM. Violin plots depict the density of individual participants' accuracy scores, with wider sections indicating higher density. Individual participant scores are provided as a scatterplot to the left of the box plots.

Critically, we also found the predicted interaction between strategy and sequence, $F(1, 288) = 20.39$, $p < 0.001$, $\eta_p^2 = 0.07$. Specifically, when instructed to find a rule, participants classified items correctly based on the rule significantly more often in the blocked condition ($M = 0.57$, $SD = 0.43$) than in the interleaved condition ($M = 0.26$, $SD = 0.37$), $t(130) = 4.40$, $p < 0.001$, $d = 0.76$. When instructed to memorize, participants classified items correctly based on similarity numerically—but not significantly—more often in the interleaved condition ($M = 0.73$. $SD = 0.35$) than in the blocked condition ($M = 0.64$, $SD = 0.32$), $t(157) = 1.61$, $p = 0.109$, $d = 0.26$.[2]

3.2.3. Performance Across Delay

The analysis of performance across delay uses only those participants who returned for Session 2 and were not excluded for data integrity reasons ($n = 228$).

Figure 3 shows correct transfer performance based on instructions for those instructed to find a rule (top panel) and memorize (bottom panel) as a function of sequence at all three time points (Session 1, Session 2—beginning, Session 2—end). Figure 3 shows that performance for those instructed to memorize—but not for those instructed to find a rule—dropped from Session 1 to the beginning of Session 2, but rebounded by the end of

Session 2. With the exception of the apparent timing effects, it appears that the interaction shown in Session 1 persisted at the beginning of Session 2 and at the end of Session 2.

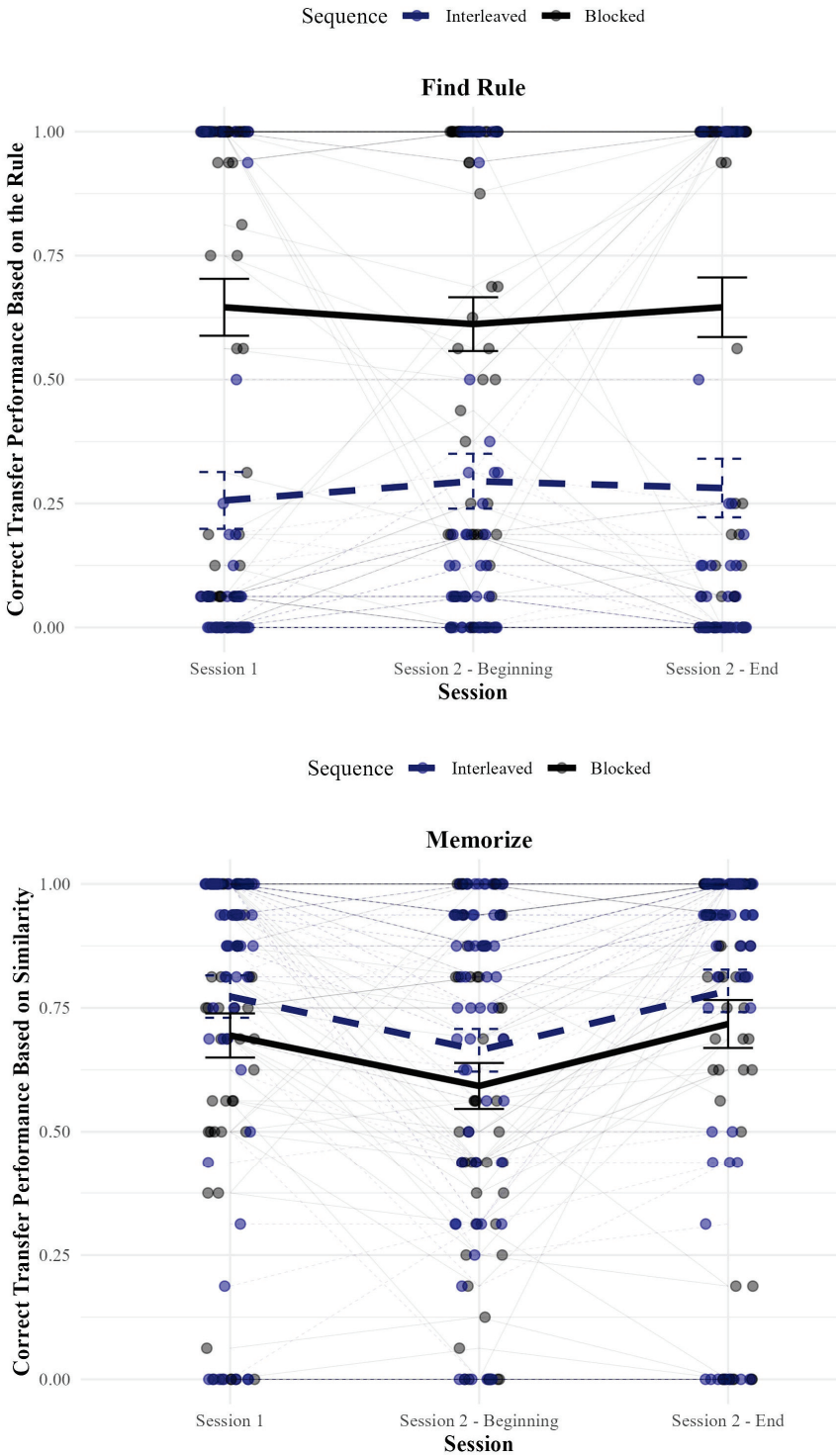


Figure 3. Correct classification based on instructions as a function of strategy instructions and sequence at the end of Session 1, beginning of Session 2, and end of Session 2. The top panel shows transfer performance for those instructed to find a rule. The bottom panel shows transfer performance for those instructed to memorize. Individual participant scores are provided as scatterplots, and performance for a given individual across time is marked with a line. Error bars represent =/− 1SEM. 95% confidence intervals.

To assess these apparent effects, we conducted a 2 (strategy: find rule vs. memorize) × 2 (sequence: interleaved vs. blocked) × 3 (time: Session 1, Session 2—beginning, Session 2—end) ANOVA on transfer performance. First, we will describe the main effects. There was a main effect of strategy such that those instructed to memorize classified items based on similarity did better (M = 0.71, SE = 0.03) than did those instructed to find a rule classify items based on the rule (M = 0.45, SE = 0.04), $F(1, 224) = 20.21$, $p < 0.001$, $\eta_p^2 = 0.08$. There was a main effect of sequence such that blocking was better (M = 0.65, SE = 0.03) than interleaving (M = 0.51, SE = 0.03) for performance, $F(1, 224) = 8.61$, $p = 0.004$, $\eta_p^2 = 0.04$. Finally, there was a main effect of time, $F(2, 436) = 11.12$, $p < 0.001$, $\eta_p^2 = 0.05$. Specifically, performance dropped from Session 1 (M = 0.59, SE = 0.03) to the beginning of Session 2 (M = 0.54, SE = 0.03), but then rebounded by the end of Session 2 (M = 0.61, SE = 0.03). Post-hoc LSD tests confirmed that there was a significant difference between Session 1 and the beginning of Session 2 ($p < 0.001$) and between the beginning of Session 2 and the end of Session 2 ($p < 0.001$), but no difference between Session 1 and the end of Session 2 ($p = 0.284$).

Next, we discuss interactions. Although there was the expected two-way interaction between strategy and sequence, $F(1, 224) = 20.21$, $p < 0.001$, $\eta_p^2 = 0.08$, there was not a three-way interaction between strategy, sequence, and time, $F(2, 448) = 0.99$, $p = 0.374$, $\eta_p^2 = 0.004$, indicating that the interaction between strategy and sequence persisted over the delay and re-training period. The average test performance across all three tests was examined using paired comparisons for interleaving versus blocking for those instructed to find a rule and those instructed to memorize. Looking at those average scores, for those instructed to find a rule, blocking was significantly more effective (M = 0.63, SD = 0.41) than interleaving (M = 28, SD = 0.40), $t(105) = 4.54$, $p < 0.001$, $d = 0.88$. For those instructed to memorize, interleaving was numerically—albeit not significantly—more effective (M = 0.75, SD = 0.31) than blocking (M = 0.67, SD = 0.32), $t(119) = 1.30$, $p = 0.196$, $d = 0.24$.

There was also an interaction between time and strategy, $F(2, 448) = 8.91$, $p < 0.001$, $\eta_p^2 = 0.04$, suggesting that those instructed to memorize experienced a drop in performance between the first test and the second (which rebounded after re-training), but performance for those instructed to find a rule did not change across time. A one-way repeated-measures ANOVA confirmed that those instructed to memorize showed different performance over time, $F(2, 240) = 22.04$, $p < 0.001$; $\eta_p^2 = 0.16$; LSD tests revealed that performance went down from Session 1 to the beginning of Session 2, $p < 0.001$, and then up from the beginning of Session 2 to the end of Session 2, $p < 0.001$. There was no effect of time on performance for those instructed to find a rule, however—$F(2, 212) = 0.84$, $p = 0.841$; $\eta_p^2 = 0.002$.[3] The interaction between time and sequence was not reliable, $F(2, 448) = 0.80$, $p = 0.451$, $\eta_p^2 = 0.004$.

3.2.4. Considering Both Experiments

Overall, we replicated the basic pattern of results shown by Little et al. (2025). Specifically, in both experiments, we found an interaction between strategy and sequence such that for those instructed to find a rule, blocking tended to be more effective than interleaving, but for those instructed to memorize, interleaving tended to be more effective than blocking. However, the exact pattern of the results differed a bit between the two studies. In the present Experiment 1, we obtained a significant effect for those instructed to memorize, but not for those instructed to find a rule, and in Experiment 2, we saw the reverse (although the effect was present for those instructed to memorize in the non-parametric comparison). We will return to this inconsistency in the General Discussion. Because the training and test procedures were exactly the same for the studies on the first day, we collapsed over both studies to examine performance for trained and transfer items. A 2 (strategy: find rule vs.

memorize) × 2 (sequence: interleaved vs. blocked) ANOVA revealed a significant effect of interleaving (M = 0.86, SE = 0.02) over blocking (M = 0.80, SE = 0.02) on trained items, $F(1, 455) = 7.36$, $p = 0.007$, $\eta_p^2 = 0.02$. There was no main effect of strategy, $F(1, 455) = 2.00$, $p = 0.159$, $\eta_p^2 = 0.004$, nor was there an interaction, $F(1, 455) = 1.70$, $p = 0.193$, $\eta_p^2 = 0.004$.

The pattern of results for transfer items reveals a significant interaction of strategy and sequence, $F(1, 455) = 28.53$, $p < 0.001$, $\eta_p^2 = 0.06$. For those instructed to find a rule, blocking was significantly more effective (M = 0.53, SD = 0.43) than interleaving (M = 0.29, SD = 0.39), $t(213) = 4.34$, $p < 0.001$, $d = 0.59$, and for those instructed to memorize, interleaving was significantly more effective (M = 0.74, SD = 0.33) than blocking (M = 0.60, SD = 0.34), $t(241) = 3.07$, $p = 0.002$, $d = 0.39$.[4]

3.2.5. Discussion

The present experiment examined transfer performance as a function of strategy (find rule vs. memorize), sequence (interleaved vs. blocked), and time (Session 1—end vs. Session 2—beginning vs. Session 2—end). At the end of Session 1, we found an interaction between strategy and sequence such that when participants were instructed to find a rule, they performed better in classifying items based on the rule when they had seen those items blocked than when they had seen them interleaved. However, when they were instructed to memorize, they performed numerically better when they received items in an interleaved versus in a blocked sequence (the difference was significant when assessing with non-parametric tests, which are more robust to the violations of normality in our data). This interaction pattern persisted across time, with performance dropping for those instructed to memorize—but not for those instructed to find a rule—between the end of Session 1 and the beginning of Session 2, and then rebounding by the end of Session 2. Finding that performance did not drop for those instructed to find a rule is consistent with the findings by Dominowski and Buyer (2000), who showed that people remember their solutions to problems, presumably because figuring out the solution to a problem is memorable (and possibly involves a deeper level of processing, Craik & Lockhart, 1972; c.f., Jacoby, 1978). Taken together, these results suggest that (a) the relative advantage of a given sequence on transfer does not change much over a delay, and (b) even when given a chance to re-train, participants persist in the same strategies.

The pattern of results at the end of Session 1 was reasonably consistent with that shown by Little et al. (2025) in their Experiment 2. However, performance—in general—was better in the present study than in their Experiment 2, especially for those participants who were instructed to find a rule. The pattern of results mostly persisted over time, which is consistent with the findings of other long-term retention and transfer research (e.g., Rohrer et al., 2015; Zulkiply, 2013).

When considering the results from both experiments (on the first day), we found that interleaving was better than blocking for trained items. This is consistent with the basic spacing effect (Cepeda et al., 2006). For transfer items, the optimal sequence depended on the instructed strategy. For those instructed to find a rule, blocking was more effective than interleaving. For those instructed to memorize, interleaving was more effective than blocking.

4. General Discussion

Across two experiments, we found an interaction effect between strategy and sequence on learning. Specifically, for those instructed to find a rule, blocking tended to be more effective than interleaving, but for those instructed to memorize, interleaving tended to be more effective than blocking. The present experiment 1 mostly replicated the results of an

earlier study (Little et al., 2025, Exp. 2), and the present Experiment 2 extended the results, showing that the pattern of results persists over a delay and even after a re-training phase.

4.1. Optimal Sequence for Learning Depends on Strategy

The benefit of interleaving over blocking for those instructed to memorize was large in Experiment 1 ($M_{difference}$ = 0.22, d = 0.65) and relatively small in Experiment 2 ($M_{difference}$ = 0.09, d = 0.26). Similarly, for those instructed to find a rule, the benefit of blocking over interleaving was relatively small in Experiment 1 ($M_{difference}$ = 0.14, d = 0.33) but large in Experiment 2 ($M_{difference}$ = 0.31, d = 0.76). There was no difference in the methods at the point at which participants took the test, so to what extent can we trust these results? The differences between interleaving and blocking in the present Experiments 1 and 2 are well within the 95% confidence intervals for differences in means reported by Little et al. (2025) for their Experiment 2, which used the same materials and procedure (for those instructed to memorize, 95% $CI_{difference}$ [0.05, 0.30]; for those instructed to find a rule, 95% $CI_{difference}$ [0.06, 0.36]). What is clear across both of the present studies is that the interaction is reliable, with a large effect. This interaction between strategy and sequence has been demonstrated in the two studies here as well as in Little et al.'s (2025) Experiment 2. Additionally, Little et al. (2025) reported two experiments examining optimal sequence as a function of self-chosen strategy, and both of those studies showed a similar pattern of results (with those findings having implications for the role of individual differences in strategy, a point we return to later).

The pattern of results for those instructed to memorize is mostly consistent with research showing that interleaving tends to be more effective than blocking when tests involve similarity-based classification that relies on memory (e.g., painting styles, Kornell & Bjork, 2008; physics problems, Samani & Pan, 2021). With regard to the performance over a delay, the findings are consistent with past research that has shown that the benefit of interleaving over blocking tends to persist from immediate to delayed tests (Firth et al., 2021). A point worth mentioning is that in the present research, it is not clear whether interleaving—that is, mixing up items from different categories—is what provides a memory advantage. In the present study, the interleaved condition also included spaced repetitions because each item was shown six times (plus once as an attention check). Under these circumstances, the spacing effect might explain enhanced performance for similarity-based classification (Cepeda et al., 2006). Indeed, when collapsing across both studies, we obtained a benefit of interleaving over blocking for trained items as well, evidencing a spacing effect for trained items. That is, had we only presented each item once, it is unclear whether there would have been any advantage of interleaving. Also worth mentioning here is that although the transfer test for those instructed to find a rule was clearly a test of transfer (not just memory), the "transfer" test for those instructed to memorize was more like a basic memory test, being a *very* near transfer test.

The pattern of results found here for those instructed to find a rule is consistent with research by Noh et al. (2016), who showed that when one needs to abstract an articulable rule, blocking is more effective than interleaving. With regard to the performance over a delay, the findings that rule-based learning is resistant to forgetting and maintains an advantage of blocking over interleaving at a delay are novel. This effect has not been shown in the category-learning literature. These results are consistent, however, with the idea that once a learner has solved a problem, their memory for the solution receives some protection from forgetting (Dominowski & Buyer, 2000). Specifically, Dominowksi and Buyer showed that learners who solved a variety of published problems (e.g., horse trade, farmer, water jars, nine-dot) had almost perfect memory for the solutions one week later. Interestingly, Dominowski and Buyer compared memory for "solvers" to memory for participants who

were exposed to the answer, and memory for solvers was far better. Although there is not much research on the retention of category rules over time, Dominowksi and Buyer's results suggest that it would depend on whether the learner would have to figure out the rule themselves (i.e., solve the problem) or if the rule was given to them (c.f., Jacoby, 1978). In fact, in some research in which items are defined by a rule, interleaving outperforms blocking, but in most cases, these are situations in which the rule is given to participants (e.g., math formulas, Rohrer & Taylor, 2007). It would be interesting to see if such rule-based learning was resistant to forgetting.

4.2. A Role for Individual Differences

Previous research has revealed individual differences in propensities to find rules versus memorize them in a variety of cognitive tasks (e.g., category learning, e.g., Craig & Lewandowsky, 2012; causal learning, e.g., Zaman et al., 2023; function learning, McDaniel et al., 2014; multiple-cue judgements, e.g., Juslin et al., 2003; skill acquisition, e.g., Bourne et al., 2010; see Little & McDaniel, 2015b for a review), including category learning studies with materials similar to the ones used here (Goldwater et al., 2018; Gouravajhala et al., 2020; Little & McDaniel, 2015a; Little et al., 2025; Liu et al., 2023; Wahlheim et al., 2016). In fact, Little et al. (2025, Exps. 1a/b) showed that when given general learning instructions (i.e., all participants were instructed "to learn", without the mention of memorization or rule finding), participants chose their own strategy, with some reporting that they oriented towards memorization and others reporting that they oriented towards rule-abstraction (see also Goldwater et al., 2018; Little & McDaniel, 2015a). Little et al.'s (2025) pattern of results when looking at self-selected strategies mimicked the pattern shown in the present experiments. For memorizers, interleaving tended to outperform blocking, but for rule-abstractors, blocking tended to outperform interleaving.

In the present research, some learners instructed to memorize found rules, and some learners instructed to find rules memorized. Figures 1–3 all show many cases in which participants who were told to find the rule failed to classify any items correctly, indicating that they classified all of the items based on similarity. Similarly, those figures also all show many cases in which participants who were told to memorize failed to classify any items correctly based on similarity, indicating that they classified all of the items correctly based on the rule.

So, to what extent can participants be instructed to use a given strategy? The manipulation check for Experiment 1 failed to reveal a significant effect of instructions on strategy ratings. However, the manipulation check in Experiment 2 was reliable (as were the correlations between instructed strategy and rule-based and similarity-based transfer performance in both Experiments 1 and 2). Nevertheless, the relatively small effects of instruction on ratings suggest that some participants may have used their predisposed strategy in spite of the instructions, perhaps because they believed that their strategy would be more effective for the task. It is also possible that participants tried to use the instructed strategy but changed their strategy (e.g., they were told to memorize but they noticed the rule; were told to find the rule but reverted to memorizing when they could not find it). Of course, it is also possible that participants failed to read the instructions. This last concern was partially mitigated by asking participants to report the instructions before the task and providing feedback (i.e., again instructing them to either memorize or find a rule). It is prudent to consider the role of multiple factors (e.g., task demands, individual propensities, expectations) in how a learner approaches a task (c.f., McDaniel & Butler, 2011).

4.3. Additional Limitations

Often, in studies examining long-term retention, participants receive a test only at the long-term retention period (or delay is manipulated between subjects). In order to maximize power, we opted to make delay a within-subjects variable, similar to the procedure used by Carvalho and Goldstone (2014a). Whether the results would be different had participants not taken a transfer test at the end of Session 1 is an open question. It is possible that the transfer test solidified participants' learning, especially for the rule condition, and had they not had that test, performance would have been much worse in the rule-based condition after a delay (c.f. testing effects, Roediger & Karpicke, 2006). Inconsistent with Carvalho and Goldstone, however, we tested transfer performance twice during the second session: once before re-training and once after. We did this because we wanted to measure the pure retention of learning, but we also wanted to allow for the possibility that learners—particularly those in the rule condition—might remember the general rule but mix up the specific letter–category combinations. Interestingly, we found that they tended to retain the general rule and the specifics, negating our need to analyze re-training performance.

A related question is why we elected to give all participants an interleaved schedule during re-training. First, in the case that performance dropped for rule finders, we wanted to be able to assess how quickly their performance improved. To do so, we needed to use test–feedback trials rather than presentation trials; and when using test–feedback trials, it does not make sense to compare interleaved to blocked re-training because, for interleaved re-training, answers are far less predictable. The second reason is that we wanted re-training to serve as a reminder of past learning, not supplant it. In particular, we did not want people to learn the rule during the second session, which would have been more likely to happen if re-training was blocked. If the reader feels that using interleaving re-training complicates the results, they can consider just the results at the end of Session 1 and at the beginning of Session 2. Considering only the first two transfer tests, the results provide the same interpretation we have put forth so far—Strategy and sequence interact to affect learning, and this pattern does not change over a delay. Additionally, similarity-based performance is lower after a delay, but rule-based performance is not.

Those instructed to find a rule tended to perform worse in classifying based on the rule than did those instructed to memorize perform at classifying based on similarity. Although our goal was not to compare across this variable, this finding suggests that the rule condition entailed a harder task than did the memorize condition. We think that this is because finding the rule—particularly in the interleaved condition—is a harder task than is memorizing, at least in the present context. One might argue that one reason that the memorize task is easier is that learners can correctly classify based on similarity with only partial memory of the trained stimuli (e.g., some of the letters from quitinoct, with the exception of the last letter). Those instructed to find a rule, on the other hand, would need memory of the last letter for each category (and either to have figured out the rule or forgotten all of the other letters); without that knowledge, performance would be poor. Although not necessarily a problem for the current experiments (because our goal was not to compare memorization performance to rule-based performance), it does have some implications for educational contexts, a point we will address in the next section.

Another limitation is that this study only uses one set of materials, and these results may not generalize to other rule-based categories. Specifically, the rule that "last letter determines category membership" is relatively easy to remember, and the specific letters may not have been that much harder to remember. Thus, the "problem-solving" nature of the task is relatively high, but the need for memory to retain the rule is relatively low. We wager that a rule with higher memory demands may not show the same pattern of results,

and would likely result in decreased performance over time. Additionally, transfer items that look similar to trained items but that should be classified differently based on a rule may be uncommon in educational contexts (although see Little & McDaniel, 2015b; Little et al., 2025 for examples; and we return to this issue in the next section).

Finally, although we framed this paper as research with application to education, and although we used an educationally realistic delay between learning and test in Experiment 2, the materials used in the present paper lack external validity to educational contexts. Nevertheless, the current results suggest that future research should be conducted with educational materials, and in the next section, we outline possibilities.

4.4. Educational Applications

Interestingly, education is often set up to promote blocked learning. For example, textbooks are often ordered in blocks of categorical information, with the idea that you must master one category of information before moving on to the next. But an advantage of interleaving has been shown across a wide array of materials, including painting style (Kornell & Bjork, 2008), naturalistic categories (Wahlheim et al., 2011), math (Rohrer & Taylor, 2007; Samani & Pan, 2021), and even simple motor skills (Wulf & Shea, 2002). Research supporting the advantage of interleaving over blocking over such a wide range of materials would suggest that a blocked sequence is always worse. Our results (and others, e.g., Noh et al., 2016) show that is not the case.

So when might blocking be advantageous? We describe two examples (concepts in psychology, concepts in analogical reasoning). These examples have, to our knowledge, not been empirically tested. As such, we do not know whether the results would align with the ideas we have put forth here, but we think that they would, and we provide them as possible future directions to infer the ecological validity of these results for education.

In the classroom, there are sometimes situations in which an instructor provides several examples of a concept, with those examples lacking superficial similarity but meant to convey an overarching idea. For example, instructors often teach about encoding specificity (i.e., retrieval success depends on the overlap between cues available at encoding and retrieval, Tulving & Thomson, 1973) in introductory psychology or cognitive psychology courses. In doing so, they tend to provide experiments from a variety of situations, including, for example, Godden and Baddeley's (1975) research examining context-dependent memory as a function of environment, Eich and Metcalfe's (1989) research examining context-dependent memory as a function of mood, and Fisher and Craik's (1977) research examining memory as a function of the match between encoding tasks and transfer tasks. Although students sometimes struggle to learn this concept even when related experiments are presented in a blocked order, we wager that they would fare even worse at abstracting the rule (i.e., what encoding specificity is/how to apply it) if the specific examples were presented far apart and interleaved with other memory concepts (which should promote item-specific memory). Based on the problem-solving literature (Dominowski & Buyer, 2000), it is worth considering that having students try to figure out the connection between these research experiments—although very difficult—might help them to retain the concept (analogous to figuring out the rule in the current research), but this is a question for future research. In the experience of the first author, students often understand the connection between the first two studies (both context effects), but they have a difficult time connecting Fisher and Craik's study to encoding specificity, instead simply linking it with the idea of levels of processing (i.e., the idea that semantic processing of a stimulus leads to better memory of it than does processing based on sounds or physical characteristics; Craik & Lockhart, 1972), owing to the similarity of the research in those two examples (and one of the authors' names).

For this encoding specificity example, a test question analogous to the transfer questions used in the present experiments might look like a levels-of-processing question (e.g., what orienting task should lead to the best performance?) but ask about predictions—based on the encoding specificity principle—for performance on a phonological-type (i.e., sound-based) test. The choices would be orienting tasks that were not explicitly discussed in class. A student who did not fully understand the idea of encoding specificity would likely predict that a semantic orienting task is better, whereas a learner who had understood the concept would choose an orienting task that is more phonological in nature.

The idea that learning often depends on abstracting a (possibly hard to notice) deep structure (or features) while ignoring superficial structure (or features) is a prominent idea in research on expertise (Chi et al., 1981). Experts develop a deep understanding of their topics and answer new problems based on concepts/rules rather than similarity (Chi et al., 1981). In some cases, such rules or concepts may have been instructed, but over time experts likely abstract consistencies and patterns. The juxtaposition between surface similarities and deep similarities is also discussed within the analogical reasoning literature (Gick & Holyoak, 1980, 1983), and the ideas presented here could help make clear predictions for future work with those materials (see also, Richland & McDonough, 2010).

For example, in a well-known example, Gick and Holyoak provided participants with Duncker's (1945) "radiation problem", and they were asked to come up with a solution to destroy a tumor. Prior to the problem, some participants were given a story about a general who wanted to overtake a fortress that shared structural/deep features—but not superficial features—with the tumor problem (i.e., the story introduced a convergence solution for overtaking the fortress that could be used to solve the tumor problem). Gick and Holyoak found that participants struggled in their ability to use the story about the general attacking the fortress to help solve the tumor problem unless they were explicitly given a hint that the two were related. However, in another study, Gick and Holyoak (1983) found that participants were increasingly likely to be able to solve the problem—without a hint—if they had received two stories with the same structural features (i.e., convergence solution) but different superficial features (e.g., one about the general and the fortress and another about a fire chief and a fire). Two superficially dissimilar stories were better than two superficially similar stories. Providing both stories allowed participants to figure out the commonality that actually mattered—that is, the convergence principle. This structural similarity was the "diagnostic feature that did not jump out at them". Although Gick and Holyoak (1980, 1983) did not examine interleaving versus blocking with their materials, we wager that participants would be more likely to abstract the underlying principle/deep structure if stories with the same rule were blocked rather than interleaved with stories featuring a different principle. One might imagine a transfer question that shared superficial similarities with one study (fortress), but had a problem that necessitated a different (non-convergence) solution that was provided within another set of materials.

5. Conclusions

In conclusion, in the present studies, we showed that the optimal sequence can depend on the strategy that participants use to learn, assessed on tests at both short and long delays. When memorizing, interleaving tends to be more effective than blocking; when trying to find a rule, blocking tends to be more effective than interleaving. This pattern of results persists over a delay, and even after a re-training phase, showing evidence that rule-based knowledge may be more resistant to forgetting than memory-based knowledge. Based on the present results, we suggest that future research explore these ideas with more educationally realistic materials.

Author Contributions: Conceptualization, J.L.L. and J.A.N.; methodology, J.L.L. and J.A.N.; software, J.L.L.; validation, J.L.L.; formal analysis, J.L.L.; investigation, J.L.L. and J.A.N.; resources, J.L.L.; data curation, J.L.L.; writing—original draft preparation, J.L.L.; writing—review and editing, J.L.L. and J.A.N.; visualization, J.A.N.; supervision, J.L.L.; project administration, J.L.L.; funding acquisition, J.L.L. All authors have read and agreed to the published version of the manuscript.

Funding: Support for Experiment 2 was provided by a 2021–2022 Faculty Support Grant from the California State University East Bay Division of Academic Affairs to Jeri Little.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of California State University, East Bay (protocol code 2019-344-F, approved 29 October 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data are available at https://osf.io/z8b69/?view_only=71fa1b74f58a490a84ccc6b73f8a82d1 (data posted 28 March 2025).

Acknowledgments: We thank Jayde Holt-Wyindon for her help recruiting participants for Session 2 in Experiment 2.

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

1 Following the hypothesis, these were planned comparisons. Because the scores were not normally distributed, we also conducted independent-sample Mann–Whitney U tests to corroborate the interpretation of these planned comparisons. Indeed, for those instructed to memorize, interleaving was better than blocking, $U = 572.50$, $p = 0.005$; for those instructed to find a rule, the difference was unreliable, $U = 1024.50$, $p = 0.113$.

2 We also conducted independent-samples Mann–Whitney U tests to corroborate the interpretation of these planned comparisons. Indeed, for those instructed to find a rule, the difference was reliable, $U = 3179.50$, $p < 0.001$. Additionally, the non-parametic test revealed that for those instructed to memorize, interleaving was better than blocking, $U = 2581.50$, $p = 0.043$.

3 Because we did not see any drop in performance for rule-finders, we did not analyze re-training performance.

4 We also conducted independent-samples Mann–Whitney U tests to corroborate the interpretation of these planned comparisons. Indeed, for those instructed to find a rule, blocking was more effective than interleaving, $U = 7810.00$, $p < 0.001$. For those instructed to memorize, interleaving was more effective than blocking, $U = 5708.00$, $p = 0.001$.

References

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, *41*, 392–402. [CrossRef]

Bourne, L. E., Jr., Raymond, W. D., & Healy, A. F. (2010). Strategy selection and use during classification skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 500–514. [CrossRef] [PubMed]

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. [CrossRef]

Carvalho, P. F., & Goldstone, R. L. (2014a). Effects of interleaved and blocked study on delayed test of category learning generalisation. *Frontiers in Psychology*, *5*, 936. [CrossRef] [PubMed]

Carvalho, P. F., & Goldstone, R. L. (2014b). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*(3), 481–495. [CrossRef]

Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, *22*(1), 281–288. [CrossRef]

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1699–1719. [CrossRef]

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380. [CrossRef]

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152. [CrossRef]

Craig, S., & Lewandowsky, S. (2012). Whichever way you choose to categorize, working memory helps you learn. *Quarterly Journal of Experimental Psychology*, *65*, 439–464. [CrossRef]

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior*, *11*(6), 671–684. [CrossRef]

Dobson, J. L. (2011). Effect of selected 'desirable difficulty' learning strategies on the retention of physiology information. *Advances in Physiology Education*, *35*, 378–383. [CrossRef] [PubMed]

Dominowski, R. L., & Buyer, L. S. (2000). Retention of problem solutions: The re-solution effect. *The American Journal of Psychology*, *113*(2), 249–274. [CrossRef] [PubMed]

Duncker, K. (1945). On problem solving. *Psychological Monographs*, *58*(5), i–113. [CrossRef]

Eglington, L. G., & Kang, S. H. (2017). Interleaved presentation benefits science category learning. *Journal of Applied Research in Memory and Cognition*, *6*, 475–485. [CrossRef]

Eich, E., & Metcalfe, J. (1989). Mood dependent memory for internal versus external events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(3), 443–455. [CrossRef]

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. [CrossRef]

Firth, J., Rivers, I., & Boyle, J. (2021). A systematic review of interleaving as a concept learning strategy. *Review of Education*, *9*(2), 642–684. [CrossRef]

Fisher, R. P., & Craik, F. I. M. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(6), 701–711. [CrossRef]

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355. [CrossRef]

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1–38. [CrossRef]

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331. [CrossRef]

Goldwater, M. B., Don, H. J., Krusche, M. J., & Livesey, E. J. (2018). Relational discovery in category learning. *Journal of Experimental Psychology: General*, *147*(1), 1–35. [CrossRef] [PubMed]

Gouravajhala, R., Wahlheim, C. N., & McDaniel, M. A. (2020). Individual and age differences in block-by-block dynamics of category learning strategies. *Quarterly Journal of Experimental Psychology*, *73*(4), 578–593. [CrossRef] [PubMed]

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*(6), 649–667. [CrossRef]

Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 924–941. [CrossRef]

Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97–103. [CrossRef]

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, *19*(6), 585–592. [CrossRef]

Kurtz, K. H., & Hovland, C. I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology*, *51*(4), 239–243. [CrossRef]

Linderholm, T., Dobson, J., & Yarbrough, M. B. (2016). The benefit of self-testing and interleaving for synthesizing concepts across multiple physiology texts. *Advances in Physiology Education*, *40*(3), 329–334. [CrossRef]

Little, J. L., & McDaniel, M. A. (2015a). Individual differences in category learning: Memorization versus rule abstraction. *Memory & Cognition*, *43*(2), 283–297. [CrossRef]

Little, J. L., & McDaniel, M. A. (2015b). Some learners abstract, others memorize examples: Implications for education. *Translational Issues in Psychological Science*, *1*(2), 158–169. [CrossRef]

Little, J. L., Nepangue, J. A., & Longares, A. (2025). The optimal sequence for learning can depend on one's strategy: An individual differences approach. *Learning and Individual Differences*, *120*, 102648. [CrossRef]

Liu, Z., Liao, S., & Seger, C. A. (2023). Rule and exemplar-based transfer in category learning. *Journal of Cognitive Neuroscience*, *35*(4), 628–644. [CrossRef] [PubMed]

McDaniel, M. A., & Butler, A. C. (2011). A contextual framework for understanding when difficulties are desirable. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 175–198). Psychology Press.

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*, *143*, 668–693. [CrossRef]

Noh, S. M., Yan, V. X., Bjork, R. A., & Maddox, W. T. (2016). Optimal sequencing during category learning: Testing a dual-learning systems perspective. *Cognition*, *155*, 23–29. [CrossRef]

Richland, L. E., & McDonough, I. M. (2010). Learning by analogy: Discriminating between potential analogs. *Contemporary Educational Psychology*, *35*(1), 28–43. [CrossRef]

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. [CrossRef]

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, *107*(3), 900–908. [CrossRef]

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*, 481–498. [CrossRef]

Samani, J., & Pan, S. C. (2021). Interleaved practice enhances memory and problem-solving ability in undergraduate physics. *NPJ Science of Learning*, *6*(1), 32. [CrossRef] [PubMed]

Sorensen, L. J., & Woltz, D. J. (2016). Blocking as a friend of induction in verbal category learning. *Memory & Cognition*, *44*, 1000–1013. [CrossRef]

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373. [CrossRef]

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, *39*, 750–763. [CrossRef]

Wahlheim, C. N., McDaniel, M. A., & Little, J. L. (2016). Category learning strategies in younger and older adults: Rule abstraction and memorization. *Psychology and Aging*, *31*(4), 346–357. [CrossRef]

Weisberg, R. W., & Alba, J. W. (1981). An examination of the alleged role of "fixation" in the solution of several "insight" problems. *Journal of Experimental Psychology: General*, *110*(2), 169–192. [CrossRef]

Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review*, *9*(2), 185–211. [CrossRef]

Zaman, J., Yu, K., & Lee, J. C. (2023). Individual differences in stimulus identification, rule induction, and generalization of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*(6), 1004–1017. [CrossRef]

Zulkiply, N. (2013). Effect of interleaving exemplars presented as auditory text on long-term retention in inductive learning. *Procedia—Social and Behavioral Sciences*, *97*, 238–245. [CrossRef]

Zulkiply, N., & Burt, J. S. (2013). Inductive learning: Does interleaving exemplars affect long-term retention? *Malaysian Journal of Learning and Instruction*, *10*, 133–155. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Monitoring-Based Rewards Enhance Both Learning Performance and Metacognitive Monitoring Accuracy

Shaohang Liu [1,2,*], Christopher Kent [3] and Josie Briscoe [3]

[1] Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University, Beijing 100875, China
[2] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Beijing Normal University, Beijing 100875, China
[3] School of Psychological Science, University of Bristol, Bristol BS8 1TU, UK; c.kent@bristol.ac.uk (C.K.); j.briscoe@bristol.ac.uk (J.B.)
* Correspondence: 11132022025@bnu.edu.cn

Abstract

Utilization of monetary rewards in educational settings remains contentious due to its potential adverse effects such as performance-related anxiety, metacognitive inaccuracy, and diminished intrinsic motivation. The current study developed a novel reward-based learning paradigm wherein rewards are granted based on monitoring accuracy rather than learning performance. Specifically, learners receive rewards for items that they predict they will remember and subsequently successfully remember them during the final test. Two experiments were conducted to assess the efficacy of this paradigm: Experiment 1 focused on learning Chinese medicine images, while Experiment 2 examined the transfer of math knowledge in classroom settings. The results indicated that rewarding the alignment between performance and metacognitive accuracy improved learning performance compared to both a baseline group and a group receiving performance-based rewards. Furthermore, this paradigm effectively mitigated performance-related anxiety and preserved intrinsic motivation. Overall, our findings highlight the critical role of reward-based learning design and emphasize the importance of addressing metacognitive accuracy alongside performance in educational practice.

Keywords: reward-based learning; metacognitive monitoring; performance-related anxiety; intrinsic motivation

1. Introduction

It is well documented that monetary incentives enhance task performance (e.g., Musslick et al., 2015; Umemoto & Holroyd, 2015). In educational contexts, providing students and children with monetary reward is popular to enhance their learning. However, the mechanisms underlying reward-based learning are complex. Monetary rewards can induce some positive effects, such as increasing attention and enhancing dopamine release in the hippocampus (Deci et al., 2001; Montague et al., 1996; Murayama et al., 2010), as well as some negative effects, including reducing intrinsic motivation (Deci et al., 1999), and provoking performance-related anxiety (Benware & Deci, 1984; Vansteenkiste et al., 2004; Lepper et al., 2005). Typically, the use of monetary rewards in real classroom settings is discouraged due to their potential negative effects. However, it remains an intriguing question whether a paradigm for providing rewards could be designed to avoid these

adverse outcomes. In this study, we proposed a novel approach to reward-based learning: incentivizing metacognitive monitoring accuracy rather than learning performance.

2. Incentive Rewards and Learning Performance

Although motivation is considered a core factor influencing learning and memory (Deci et al., 1991; Dweck, 1986), the role of incentive rewards in learning and memory remains mixed. Providing performance-based rewards can have a direct positive effect by activating the mesolimbic reward system, which promotes release of dopamine in the hippocampus and facilitates memory consolidation (Knowlton & Castel, 2022). However, many studies have demonstrated various negative effects of monetary rewards on learning and memory. For instance, performance-based rewards can provoke performance-related anxiety and pressure, impairing learning outcomes (e.g., Ellis & Ashbrook, 1989). When students are motivated to perform optimally, their cognitive processes can be disrupted by irrelevant thoughts, such as worrying about poor performance or doubting their abilities, which hinders deep processing of study materials (Baumeister, 1984; DeCaro et al., 2011). Additionally, studies have shown that when learning behaviors are driven by extrinsic rewards, the level of perceived anxiety triggered by performance-based rewards may lead to a more superficial learning style characterized by shallow levels of elaboration and less effortful encoding (e.g., Benware & Deci, 1984; Simons et al., 2004; Vansteenkiste et al., 2004; Lepper et al., 2005).

Aside from this, extrinsic rewards can undermine students' intrinsic motivation for learning (Deci et al., 1999). This undermining effect is evidenced by decreased satisfaction and poor learning performance when a non-reward learning activity follows a reward-based one. Additionally, studies have shown that people's motivation for choosing challenging tasks to perform may be offset by the risk of losing out potential extrinsic rewards. Murayama and Kuhbandner (2011) even found that the beneficial effect of incentive rewards only worked for the content that participants were not interested in, but disappeared for the content that they were interested in.

The mixed effects of performance-based rewards create a dilemma. On one hand, providing monetary rewards appears to be an easy-to-implement and time-efficient technique to enhance learning performance. On the other hand, monetary rewards might produce a set of negative influences. Given these considerations, an important question arises: Is there a design of reward-based learning that can enhance learning performance and also avoid negative effects of performance-based rewards?

3. Non-Performance-Based Rewards

A primary concern with using incentive rewards in educational settings is their typical dependence on students' learning or academic performance. To address this concern, decoupling rewards from performance may offer noticeable benefits (Fröber & Dreisbach, 2016, 2020). For example, Westbrook et al. (2013) reported a paradigm that measures participant's subjective cost of cognitive effort, where participants chose between low-effort tasks for smaller rewards and high-effort tasks for larger rewards. The additional compensation required to equalize these choices quantified the subjective cost of cognitive effort. The study demonstrates that cognitive effort discounting correlates with task load, performance, and individual traits like the need for cognition. For example, older adults perceived cognitive tasks as more effortful and required significantly higher compensation than younger adults, even when performance was controlled. This suggests that providing greater rewards for increased mental effort may counteract the decline in motivation (Kool et al., 2013; Shenhav et al., 2017).

Recently, Lin et al. (2024) introduced an innovative reward-based paradigm targeting to enhance individuals' willingness to choose challenging tasks to perform by providing them rewards based on the level of difficulty of the tasks they choose to performance, rather than based on their task performance. In this study, participants in a performance group received rewards according to their task performance, whereas those in an effort group received rewards solely based on the level of difficulty of the tasks they chose to perform. After the intervention phase, both groups were invited to choose a new task among a set of tasks with different levels of difficulty to perform. The results showed that effort-based rewards increased participants' willingness to choose challenging tasks to perform, and this enhancement effect even successfully transferred to subsequent tasks in which the rewards were completely removed. Interestingly, despite that the effort group was selected to perform more difficult tasks than the performance group, overall task performance did not differ between these two groups, suggesting that non-performance-based rewards (e.g., effort-based rewards) could effectively redefine the role of incentives in enhancing task performance.

In the current study, we developed a novel reward paradigm that shifts the focus from rewarding learning performance to rewarding metacognitive monitoring accuracy. Specifically, learners are required to make a judgment of learning (JOL, i.e., estimating the likelihood of remembering a study item in a later test) after studying each item. Rewards are structured such that learners receive the highest reward for successfully remembering the items they predict they will remember, a reduced reward for successfully remembering the items they predict they will forget, and a modest reward for accurately predicting the items that are forgotten. No reward is given for the items that they predict they will remember but actually forget. It is worth noting that, in this reward paradigm, learners still receive higher rewards for items they successfully remember than for the items they forget. This design is implemented to prevent the risk that some learners may choose to not study any items and predict they will forget all items in order to obtain maximum rewards (see below for detailed discussion).

4. Rationale of Monitoring-Based Rewards

We assume that monitoring-based rewards should offer noticeable benefits and address some drawbacks of performance-based rewards in three keyways.

First, individuals can benefit from making JOLs and rewarding JOL accuracy. On the one hand, previous studies showed that making JOLs can enhance learning performance, a phenomenon known as the *JOL reactivity effect* (Janes et al., 2018; Li et al., 2022, 2024; Zhao et al., 2022). Although the reactivity effect of JOL has only been found for the related word pairs, studies have noted the beneficial mechanisms of making JOLs in improving learning. For example, the *enhanced-engagement theory* suggests that the requirement of making JOLs for each item can sustain learners' attention and improve overall learning performance by preventing attentional waning (Tauber & Witherby, 2019; Tekin & Roediger, 2020; Zhao et al., 2022). The *cue-strengthening theory* posits that learners need to search for "diagnostic" cues to guide JOL formation, and the cue-searching process in turn induces deeper level-of-processing of the study items and therefore enhances learning performance (Soderstrom et al., 2015). On the other hand, monitoring-based rewards can improve learning performance by enhancing metacognitive monitoring accuracy. Studies have shown that students with higher metacognitive monitoring accuracy tend to achieve better learning performance (Metcalfe & Finn, 2008; Van Loon & Oeri, 2023), especially during self-regulated learning. This correlation is likely because students with higher metacognitive monitoring accuracy can more appropriately allocate their cognitive resources (Bjork et al.,

2013), engage in deeper processing of the learning material (Toppino & Cohen, 2010), and adaptively adjust their learning strategies (Bjork et al., 2013).

Second, monitoring-based rewards can mitigate some negative effects (e.g., performance-related anxiety) associated with performance-based rewards. It has been well-shown that performance-related anxiety can disrupt cognitive processing and impede elaborative processing (Baumeister, 1984; DeCaro et al., 2011). This anxiety can also negatively affect metacognitive monitoring (Janes et al., 2018; Halamish & Undorf, 2022; Rivers et al., 2021), and the learner may fail to effectively allocate their study resources, such as study time. We assume the monitoring-based reward can reduce the anxiety caused by performance-based reward. On the one hand, by decoupling rewards from performance, monitoring-based rewards should be able to reduce performance-related anxiety by eliminating fear of poor performance, as such rewards are not contingent upon successful retrieval (Ashcraft & Krause, 2007). Furthermore, monitoring-based rewards are expected to enhance students' sense of control, as they receive greater rewards for items they believe they have mastered, thereby alleviating performance anxiety (Bandura & Cervone, 1983; Lachman & Weaver, 1998). Additionally, studies have shown that focusing more on intrinsic improvement (i.e., what knowledge is acquired) can reduce the anxiety caused by peer pressure (DeCaro et al., 2011). Making Judgments of Learning (JOLs) encourages students to reflect on the knowledge they have learned, rather than focusing on external factors such as "how much money is lost".

Third, different from performance-related rewards, monitoring-based rewards should not undermine students' intrinsic learning motivation. According to the *cognitive evaluation theory* (CET), extrinsic rewards can diminish intrinsic motivation by undermining feelings of self-control and competence (Lepper, 1973). This theory argues that rewards for engaging in interesting activities may lead individuals to attribute their behaviors to external rewards rather than their intrinsic interests. However, by aligning rewards with metacognitive monitoring accuracy rather than test performance, students are encouraged to reflect on their actual level of knowledge mastery rather than focusing solely on how well their final test performance will be. This approach may help maintain intrinsic motivation by emphasizing the value of accurate self-assessment over mere performance outcomes.

5. Overview of the Current Study

In the current study, we aimed to explore whether aligning rewards with metacognitive monitoring accuracy rather than test performance can enhance learning outcomes and meanwhile mitigate some drawbacks associated with performance-based rewards. To investigate this, we employed four groups of participants, inducing (a) a performance reward group, where participants' final rewards were based on their test performance, (b) a monitoring reward group, where participants' final rewards were based on their metacognitive monitoring accuracy, (c) a monitoring group, where participants made item-by-item JOLs without receiving any rewards related to monitoring accuracy, and (d) a baseline group, where participants did not make JOLs and did not receive any rewards related to monitoring accuracy and test performance. The performance reward group was not asked to make judgments of learning (JOLs) because our aim was to compare the monitor reward group with the most conventional method of providing rewards—where the amount of reward is determined solely by performance.

Based on the preceding discussions, we formulated three hypotheses: (a) learning performance should be higher in the monitoring reward group than in the other three group; (b) performance-related anxiety should be lower in the monitoring reward group than in the performance reward group; and (c) intrinsic motivation to learn new information should be higher in the monitoring reward group than in the performance reward group.

6. Experiment 1

Experiment 1 was conducted in a laboratory to test the three hypotheses discussed above.

6.1. Methods

6.1.1. Participants

As no previous studies have ever explored related effects, we did not conduct power analysis based on previously reported effect size. A power analysis performed via G*Power (Faul et al., 2007) suggested a minimum sample size of 45 participants in each group to observe a moderate effect size (Cohen's f = 0.25) at 0.80 power. Due to over-recruitment, 220 undergraduate students from a Chinese university were recruited as participants. Nine participants did not return to complete the final test and their data were excluded from analyses, leaving final data from 211 participants (M_{age} = 20.1, SD = 1.6; 100 female). There were 53 participants in the baseline group, 51 in the monitoring group, 54 in the performance reward group, and 53 in the monitoring reward group.

All participants signed an agreement to participate. They obtained monetary reward after completion of the experiment. Specifically, participants from the baseline and monitoring group received 40 RMB at the end of the second-day test. For the two reward groups, participants obtained 40 RMB for their participation and a bonus dependent on their test performance (performance reward group) or metacognitive monitoring accuracy (monitoring reward group). Note that we did not control the total amount of reward across the four experimental groups, as our aim was to investigate whether the provision of additional rewards can increase learning performance, and how this design of reward-based learning can be optimized (see Table 1 for descriptive statistics of reward amount of four experimental groups). The protocol was ethically approved by the Faculty of Psychology, Beijing Normal University.

Table 1. Mean (SD) of reward amount for performance-based and monitor-based groups.

Reward Amount (RMB)	Performance-Based Reward	Monitor-Based Reward
Experiment 1	22.61 (7.27)	33.53 (8.24)
Experiment 2	9.57 (2.24)	11.33 (2.88)

6.1.2. Material

The learning materials were 40 Chinese medicine images. Each image were 400 × 400 pixels with its name presented below. The materials have been made publicly available at OSF (https://osf.io/3czqr/?view_only=e7d6ea6f9e174b06b794d660e4b1814e, accessed on 16 May 2024). All stimuli were presented via Psychopy v2023.1.2 (Peirce, 2007).

6.1.3. Procedure

The experiment consisted of two sessions: a study session and a test session. Participants were instructed that they would learn 40 Chinese medicine images and take a final test on the following day. In the performance reward group, participants were further informed that they would earn 2 RMB for each item successfully recalled in the final test. In the monitoring reward group, participants were asked to predict whether they would be able to remember the name of each Chinese medicine image after 24 h. They were told that they would earn: (a) 2 RMB for the items that they predicted they would remember and actually successfully remembered in the final test, (b) 0.5 RMB for the items that they predicted they would forget but actually remembered in the final test, (c) 1 RMB for the items that they predicted they would forget and actually forgot in the final test, and

(d) 0 RMB for the items that they predicted they would remember but actually forgot in the final test (see Table 1 for average reward obtained for the performance reward group and monitoring reward group). Note that only when participants try their best can they obtain the most reward. The purpose of setting differential amount of reward for condition (a) and (c) is to avoid the possibility that participants just skip the item.

In the monitoring group, participants were only asked to make JOLs and did not receive any instructions about monetary rewards. The baseline group did not receive any instructions about monetary rewards and JOLs, and was solely asked to learn and memorize the image–name pairs.

During the study session, participants learned the image–name pairs one-by-one in a random order. Each pair was presented for 10 s with a 1-s interstimulus interval inserted between each pair of study items. The image was displayed at the center of the screen with its name shown beneath it. In the baseline and performance reward groups, participants viewed the image–name pair for 10 s without making JOLs. In the monitoring and monitoring reward groups, a JOL question appeared beneath the image–name pair asking, "*Will you be able to remember the name of this image after 24 h?*" Participants indicated their metacognitive judgments by clicking on either the "*WILL*" or "*WILL NOT*" button. If no response was made during 10 s, an exclamation mark reminder appeared during the interstimulus interval after the image trial.

After studying all image–name pairs, participants rated their anxiety level using a 9-point Likert scale ("*Please rate the anxiety level during learning the images. 1 = not anxious at all*; *9 = very anxious*"). Then participants were dismissed and invited to come back for the final test 24 h later. In the final test administered 24 h later, the 40 medicine images were presented one-by-one in a random order, and participants were asked to recall the name of each medicine image displayed. There was no time pressure and no feedback in the final test.

Following the final test, participants were invited to participate in an additional task involving learning another 40 Chinese medicine images using the same study-test procedure but without any rewards. Specifically, they were told that "*There will be another chance for you to learn additional 40 images with the same procedure, and you will not get any monetary rewards for this learning task. Please rate your willingness to participate in this task, and our experimenter would contact you if you are willing to participate*." They were asked to rate their willingness on a 9-point Likert scale (*1 = not willing at all*; *9 = very willing*). This question was implemented to measure participants' intrinsic motivation, and explore whether monitoring-based rewards can avoid reducing intrinsic motivation induced by performance-based rewards.

6.2. Results

A one-way ANOVA was conducted to compare the memory performance of four experimental groups. Significant differences were detected: $F(3,207) = 19.63$, $p < 0.001$, $\eta_p^2 = 0.22$, $BF_{10} > 1000$ (see Figure 1a). Post hoc comparisons, using Bonferroni corrected *t*-tests, demonstrated that the monitoring reward group ($M = 17.26$, $SD = 4.06$) achieved best test performance, and test performance in this group was significantly better than that in the monitoring group ($M = 14.04$, $SD = 3.00$), difference = 3.23 [1.30, 5.15], $t = 4.34$, $p < 0.001$, $d = 0.851$, $BF_{10} > 1000$, higher than that in the performance reward group ($M = 13.11$, $SD = 4.12$), difference = 4.15 [2.21, 6.09], $t = 5.42$, $p = 0.001$, $d = 1.092$, $BF_{10} > 1000$, and higher than that in the baseline group ($M = 11.84$, $SD = 3.83$), difference = 5.42 [3.50, 7.35], $t = 7.05$, $p < 0.001$, $d = 1.43$, $BF_{10} > 1000$. Additionally, the monitoring group successfully recalled more items compared with the baseline group, difference = 2.20 [0.25, 4.14], $t = 2.93$, $p = 0.02$, $d = 0.58$, $BF_{10} = 18.81$. There was no detectable difference in final test performance between

the baseline and performance reward groups, difference = 1.26 [−0.64, 3.16], t = 1.72, p = 0.52, d = 0.33, BF_{10} = 0.68, and no detectable difference between the performance reward and monitoring groups, difference = 0.93 [−0.97, 2.83], t = 1.27, p = 0.21, d = 0.25, BF_{10} = 0.45.

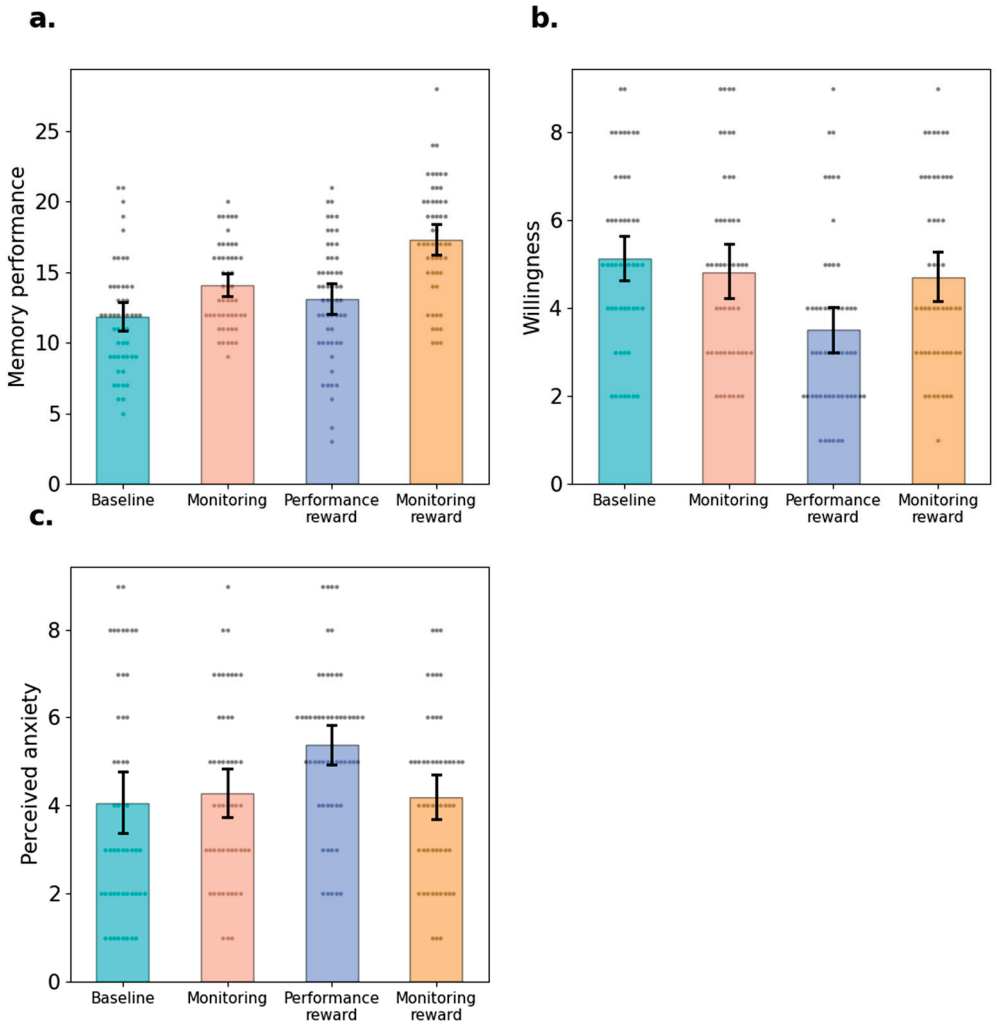


Figure 1. Results of Experiment 1. *Note.* Memory performance (Panel **a**), willingness to participate in the subsequent task without reward (Panel **b**) and test anxiety (Panel **c**) of four groups. Error bars indicate 95% confidence intervals and the grey dots represent the value of individual participants.

As shown in Figure 1b, the ANOVA shows that the four groups reported differential motivations to participate in the subsequent task in which rewards were no longer provided, $F(3,207)$ = 6.57, p < 0.001, η_p^2 = 0.09, BF_{10} = 78.98. Bonferroni corrected t-tests showed that participants in the performance reward group (M = 3.52, SD = 1.99) reported lower motivation than the baseline group, difference = −1.64 [−2.67, −0.60], t = −4.10, p < 0.001, d = −0.79, BF_{10} = 418.84, lower than the monitoring group, difference = −1.32 [−2.36, −0.29], t = −3.31, p = 0.006, d = −0.64, BF_{10} = 25.34. and lower than the monitoring reward group difference = 1.20 [−2.22, −0.17], t= −3.03, p = 0.015, d = −0.58, BF_{10} = 12.06. However, there were no differences in motivation between the baseline and the monitoring group, difference = 0.31 [−0.75,1.37], t = 0.77, p = 0.89, d = 0.15, BF_{10} = 0.27, and between the monitoring group and monitoring reward group, difference = 0.13 [−0.92, 1.17], t = 0.31, p = 0.89, d = 0.06, BF_{10} = 0.22.

As is shown in Figure 1c, the ANOVA shows that the levels of anxiety experienced during the learning phase differed significantly among the four groups, $F(3,207)$ = 4.69, p = 0.003, η_p^2 = 0.06, BF_{10} = 7.85. Specifically, Bonferroni corrected t-tests showed that

the performance reward group (M = 5.38, SD = 1.82) reported higher levels of anxiety compared with the baseline group (M = 4.04, SD = 2.57), difference = 1.34 [0.29, 2.38], t = 3.30, p = 0.007, d = 0.64, BF_{10} = 14.48, higher than the Monitoring group (M = 4.28, SD = 2.05), difference = 1.10 [0.05, 2.15], t = 2.58, p = 0.04, d = 0.53, BF_{10} = 9.05, and higher than the monitoring reward group (M = 4.19, SD = 1.87), difference = 1.19 [0.15, 2.22], t = 2.82, p = 0.02, d = 0.57, BF_{10} = 27.29. However, no difference was found between the baseline and the monitoring groups, difference = 0.24 [−0.84,1.31], t = 0.57, p = 1.00, d = 0.11, BF_{10} = 0.24., and between the monitoring and monitoring reward groups, difference = 0.09 [−0.98, 1.15], t = 0.21, p = 1.00, d = 0.04, BF_{10} = 0.21.

Metacognitive monitoring accuracy for each participant in the monitoring and monitoring reward groups was measured as the Gamma (G) correlation between predicted performance and actual performance. Metacognitive monitoring accuracy in the monitoring reward group (M of G = 0.32, SD = 0.68) was significantly higher than that in monitoring group (M = 0.02, SD = 0.66), difference = 0.30 [0.04, 0.56], $t(102)$ = 2.30, p = 0.02, d = 0.45, BF_{10} = 2.14, suggesting that monetary rewards successfully enhanced monitoring accuracy.

6.3. Discussion

Experiment 1 confirmed the effectiveness of monitoring-based rewards, demonstrating its superior benefits compared to other learning strategies or interventions, such as monitoring-only and traditional performance-based rewards. Additionally, monitoring-based rewards do not undermine intrinsic motivation and do not provoke performance-related anxiety as performance-related rewards do.

7. Experiment 2

Experiment 2 aimed to evaluate the effectiveness of monitoring-based rewards in real educational settings and assess its transferability to different learning tasks. To achieve this, Experiment 2 employed a different participant sample (high school students) and used a different type of learning materials (mathematical theorems).

7.1. Method

Participants

The sample size was not pre-determined. Instead, it was determined by the number of available classes and students in each class. In total, 246 grade 10 (high school) students from 4 classes in a high school were recruited initially, while 6 students did not attend the second-day task and hence were excluded, leaving final data from 240 students (M_{age} = 16.3, SD = 0.51; 129 female). Students from four classes were randomly assigned to four experimental groups. The size of each group was determined by the size of the class, so there were 60 students in each class and literally in each experimental group.

The academic performance of four classes (grades taken from students' mid-term exam) were well matched. Likewise Experiment 1, all participants obtained 40 RMB for participation at the end of the experiment, and two reward groups obtained a bonus dependent on their test performance (performance reward group) or metacognitive accuracy (monitoring reward group).

7.2. Materials

The learning materials comprised 10 mathematical theorems, including concepts such as *the Cosine Theorem*, which the students had not previously encountered. These theorems were presented in a ten-page format, with each page containing one theorem along with an example of its application (see OSF https://osf.io/3czqr/?view_only=e7d6ea6f9e174b06b794d660e4b1814e for the materials, accessed on 16 May 2024).

7.3. Procedure

Experiment 2 was conducted in a classroom setting, following a similar study-test procedure as that in Experiment 1, but adapted for an educational environment. Participants were instructed that they would learn 10 mathematical theorems, each accompanied by an example, and they would take a final test the following day. In the final test, participants were required to apply the theorems to solve novel problems. As in Experiment 1, participants were informed about the reward structure specific to their experimental group (see Table 1 for average reward obtained for the performance reward group and monitoring reward group).

During the learning session, all students received learning materials with a blank cover. They began reading the materials once the "begin" instruction was given by the teacher. Unlike Experiment 1, where the study phase was experimenter-paced, the study phase in Experiment 2 was self-paced. That is, in Experiment 2, students were allowed to self-regulate their time across a total of 30 min for studying the whole set of materials. For participants in the monitoring and monitoring reward groups, they needed to answer "*Can you master this theorem in the test after 24 h?*" by ticking "*CAN*" or "*CAN'T*" on each page.

To maintain ecological validity, participants were not asked to report their anxiety levels after the learning session. On the second day, participants completed a final test consisting of ten questions, where they applied the learned theorems to solve novel problems. They were given 45 min to complete the test. After finishing the final test, participants were asked about their willingness to participate in a similar task, with the same learning procedure but without any monetary rewards. They rated their willingness using a 9-point Likert scale (1 = *not willing at all*; 9 = *very willing*).

8. Results

A one-way ANOVA was conducted to compare the test performance between four experimental groups. There was a significant difference across four groups (Figure 2a), $F(3, 236) = 8.39$, $p < 0.001$, $\eta_p^2 = 0.10$, $BF_{10} > 1000$. Specifically, Bonferroni corrected *t*-tests showed that the monitoring reward group ($M = 5.87$, $SD = 1.58$) achieved better performance compared with baseline group, difference = 0.85 [0.07, 1.64], $t = 2.80$, $p = 0.03$, $d = 0.51$, $BF_{10} = 10.11$, monitoring group, difference = 0.82 [0.03, 1.60], $t = 2.70$, $p = 0.04$, $d = 0.49$, $BF_{10} = 3.54$, and performance reward group ($M = 4.30$, $SD = 1.60$), difference = 1.57 [0.78, 2.35], $t = 0.11$, $p < 0.001$, $d = 0.94$, $BF_{10} > 1000$. However, no difference was found between monitoring group ($M = 5.05$, $SD = 1.91$) and baseline group ($M = 5.02$, $SD = 1.53$), difference = 0.03 [−0.75, 0.82], $t = 0.11$, $p = 0.91$, $d = 0.02$, $BF_{10} = 0.20$. Performance of the performance reward group was numerically worse than the baseline group, difference = −0.72 [−1.50, 0.07], $t = -2.36$, $p = 0.11$, $d = -0.43$, $BF_{10} = 3.18$, and the monitoring group, difference = −0.75 [1.54, −0.04], $t = -2.47$, $p = 0.09$, $d = -0.45$, $BF_{10} = 2.21$, but neither comparison reached significance. Yet, the Bayesian statistics indicated the differences were at least anecdotal.

As shown in Figure 2b, the ANOVA shows that the four groups reported differential motivations to participate in the subsequent task in which rewards were no longer provided, $F(3, 236) = 4.01$, $p = 0.008$, $\eta_p^2 = 0.05$, $BF_{10} = 3.09$. Bonferroni corrected *t*-tests showed that participants' willingness to participate in the subsequent task was lower in the performance reward group ($M = 3.83$, $SD = 1.82$) than in the monitoring reward group ($M = 4.97$, $SD = 2.37$), difference = −1.13 [−2.13, −0.14], $t = -2.43$, $p = 0.02$, $d = -0.54$, $BF_{10} = 8.99$, lower than in the monitoring group ($M = 4.77$, $SD = 2.27$), difference = −0.93 [−1.93, −0.06], $t = 2.46$, $p = 0.046$, $d = 0.44$, $BF_{10} = 3.08$, and lower than in the baseline group ($M = 4.98$, $SD = 1.94$), difference = −1.15 [−2.15, −0.16], $t = 2.96$, $p = 0.019$, $d = 0.546$, $BF_{10} = 27.36$. There were no detectable difference between the baseline and monitoring groups, difference = 0.22

[−0.78, 1.21], $t = 0.56$, $p = 1.00$, $d = 0.10$, $BF_{10} = 0.22$, between the baseline and monitoring reward groups, difference = 0.02 [−0.98, 1.01], $t = 0.04$, $p = 1.00$, $d = 0.01$, $BF_{10} = 0.03$, and between the monitoring and monitoring reward groups, difference = 0.20 [−1.20, 0.8], $t = 0.52$, $p = 1.00$, $d = 0.10$, $BF_{10} = 0.03$.

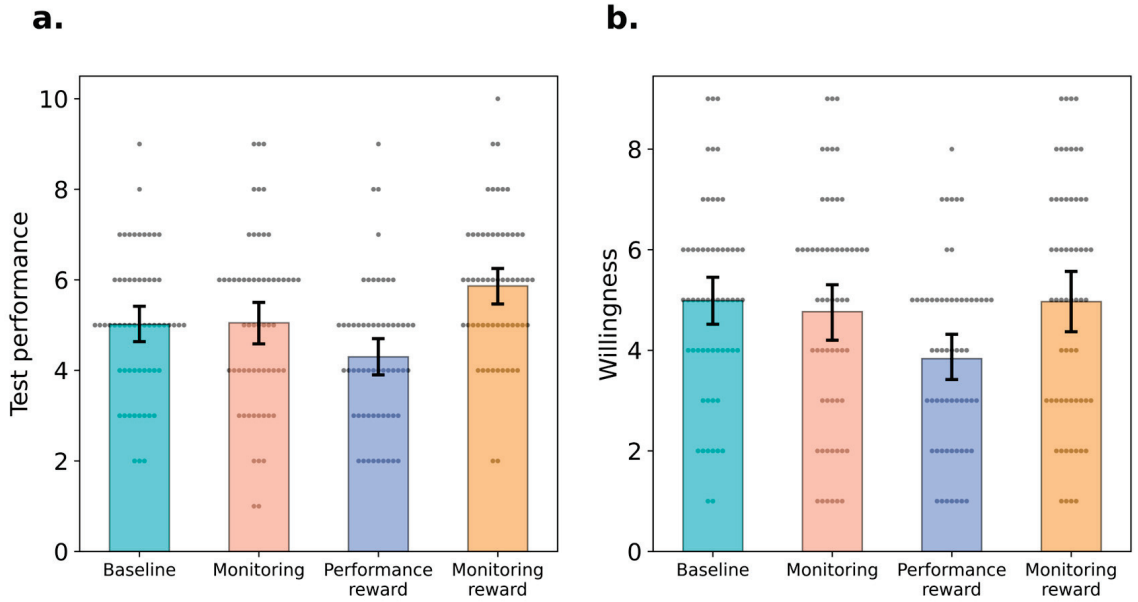


Figure 2. Results of Experiment 2. *Note.* Test performance (Panel **a**), willingness to participate in the subsequent task without reward (Panel **b**) of four groups. Error bars indicate 95% confidence interval and the grey dots represent the value of individual participants.

Likewise Experiment 1, JOL accuracy was represented by the *G* correlation between predicted performance and actual test performance. Consistent with Experiment 1, JOL accuracy in the JOL reward group ($M = 0.38$ $SD = 0.87$) was significantly greater than that in the monitoring group ($M = -0.08$, $SD = 0.74$), difference = 0.46 [0.16, 0.75], $t(115) = 3.08$, $p = 0.003$, $d = 0.57$, $BF_{10} = 12.91$.

Discussion

Consistent with Experiment 1, Experiment 2 confirmed the superior effectiveness of monitoring-based rewards over performance-based rewards. However, a notable difference was observed: providing performance-based rewards impaired test performance compared to the performance of non-reward groups. This adverse outcome may be attributed to the higher levels of anxiety associated with math learning compared to other types of learning activities (Ashcraft & Krause, 2007), which likely exacerbated the negative side effects of performance-based rewards.

9. General Discussion

In the present study, we introduced a novel paradigm for reward-based learning that emphasizes the alignment between memory performance and the accuracy of metacognitive monitoring, rather than focusing solely on memory performance. This approach led to significant improvements in learning outcomes compared to both the baseline group and traditional performance-based reward group. Moreover, we demonstrated that this effect extends beyond image-pair learning to the transfer of math knowledge, underscoring the robustness of the monitoring-based reward paradigm across various educational materials and tasks. The findings suggest that the mechanism underlying money-motivated memory—where monetary incentives activate the mesolimbic reward system and enhance

dopamine release, thereby aiding memory consolidation (see Knowlton & Castel, 2022; Miendlarzewska et al., 2016, for reviews)—remains effective, even when the criteria for earning rewards are refined. Importantly, we demonstrate that neither performance-based rewards nor the act of making JOLs alone can consistently improve learning performance as effectively as monitoring-based rewards. This finding highlights that the combination of accurate metacognitive monitoring and reward is critical for enhancing learning performance and mitigating the negative effects associated with performance-based rewards.

Studies have demonstrated that making judgments of learning (JOLs) can enhance memory performance by promoting elaborative processing and increasing attention to the learning materials (Tauber & Witherby, 2019; Tekin & Roediger, 2020; Zhao et al., 2022). Although these mechanisms were not justified in the present study, in Experiment 1, participants who made JOLs outperformed those in all no-monitoring groups (i.e., baseline and performance reward groups). Additionally, monitoring-based reward, namely associating metacognitive monitoring accuracy and the number of rewards, further improved both memory performance and metacognitive monitoring accuracy. However, in Experiment 2, making JOLs without rewards did not lead to overall performance improvements. This discrepancy may be attributed to the nature of the learning materials used—math theorems—which can make it more challenging to strengthen the cue-target association typically enhanced by JOLs. Additionally, previous studies have suggested that the reactivity effect of making JOLs is typically restricted to word-pair materials (Janes et al., 2018; Li et al., 2022, 2024; Zhao et al., 2022). Therefore, caution is warranted in attributing the benefit of monitor-based rewards solely to the act of making JOLs.

Nonetheless, the provision of rewards in Experiment 2 improved the accuracy of JOLs, potentially aiding students in adaptively adjusting their learning, particularly in self-paced tasks (for a review, see Bjork et al., 2013). Notably, in both experiments, participants demonstrated relatively low—sometimes even negative—metacognitive monitoring accuracy in the pure monitoring group when rewards were absent. However, the accuracy increased significantly with the introduction of rewards. This suggests that rewards may encourage participants to exert greater effort in their judgments of learning. Such an increase in effort not only enhances monitoring accuracy but may also promote more elaborative processing of the materials (Zhao et al., 2022), ultimately improving their learning outcomes.

Consistent with previous studies (e.g., Kuhbandner et al., 2016; Liu et al., 2022), the present study found that performance-contingent rewards did not enhance memory for image-pair learning. This outcome may be attributed to the adverse effects of performance-contingent rewards, which can offset potential benefits despite their interaction with the mesolimbic reward system and the hippocampus. The primary negative consequence of performance-contingent rewards is their tendency to induce pressure-related anxiety, leading to interfering thoughts and superficial encoding (Benware & Deci, 1984; Vansteenkiste et al., 2004; Lepper et al., 2005). Indeed, participants in the performance-reward group reported higher levels of anxiety. Additionally, performance-contingent rewards even led to poorer performance in math knowledge learning, likely because math learning can induce greater anxiety compared to other types of learning activities (Ashcraft & Krause, 2007).

The present study revealed that monitoring-based reward can mitigate performance anxiety, highlighting the positive aspects of monetary incentives. Unlike performance-based rewards, which often drive individuals to pursue extreme outcomes, monitoring-based rewards shift the focus from the fear of failure to the anticipation of potential gains. Furthermore, the compensation for items correctly predicted to be forgotten helps alleviate anxiety related to retrieval failure. Consequently, students in the monitoring-based reward group are less burdened by interfering thoughts and can allocate more cognitive resources

to elaborative processing. This is particularly important for making accurate JOLs, as the reward structure incentivizes careful and accurate metacognitive monitoring.

Another concern with providing monetary rewards is the potential undermining of students' intrinsic motivation for learning (Deci et al., 1999). This study confirmed this well-documented effect, as the performance reward group exhibited decreased motivation for continued learning when the reward was removed, whereas the motivation levels of other groups remained unchanged. Interestingly, despite the provision of monetary rewards, monitoring-based reward did not diminish students' intrinsic motivation. This may be because the decrease in intrinsic motivation was countered by a positive shift in awareness: students became more cognizant of their achievements, fostering a more positive mindset (Harackiewicz et al., 2000). Thus, the withdrawal of monetary rewards does not necessarily lead to a decline in intrinsic motivation, as the impact of monetary rewards depends on their specific application and the context in which they are used.

10. Limitation and Future Directions

One limitation of the present study is that students' anxiety and intrinsic motivation were measured using a single-item assessment after learning. Although the results from this item supported the expected hypotheses, the effects were not fully conclusive. Future studies could benefit from incorporating real-time physiological measures, such as heart rate and skin conductance, to more accurately reflect students' anxiety levels during the learning process.

Another limitation of the present study is the lack of item-level learning time data in Experiment 2. While our findings demonstrate that the benefit of monitoring rewards is robust in self-regulated learning, the specific ways in which students monitor and control their learning during this process remain unclear. Future research is needed to address this issue and provide deeper insights.

Additionally, while the effectiveness of monitoring-based reward was demonstrated, the study provided limited insight into its underlying mechanisms. Specifically, neither test anxiety levels nor JOL accuracy were found to correlate strongly with actual learning performance in the present study. This suggests that the benefits of monitoring-based reward may involve multiple components. For instance, in the case of certain materials, such as math knowledge, JOLs alone may not significantly enhance memory performance. Instead, the benefits of monitoring-based reward might stem from other factors, such as reduced anxiety or enhanced covert retrieval processes.

Several hypotheses merit further investigation in future research. First, the advantages of monitoring-based reward may diminish or even disappear with immediate recall tests. This prediction is based on two factors: (1) hippocampal consolidation typically requires time and does not occur immediately during encoding (Murayama & Kuhbandner, 2011), and (2) immediate recall might negate the benefits of covert retrieval practice, which often occurs after learning, particularly when students are aware of their knowledge mastery following explicit JOLs. Thus, monitoring-based reward may be less effective when the interval between learning and testing is minimal.

It would be valuable to investigate whether monitoring-based rewards can also enhance learning outcomes when the total learning time is limited. In both experiments conducted in the present study, the total learning time was fixed, requiring students to allocate their learning time and cognitive resources effectively (Bjork et al., 2013). This constraint may encourage students to focus on easier-to-learn items to maximize efficiency. Conversely, participants in the performance-based reward condition might allocate more time to studying difficult items in an effort to earn greater rewards. This could heighten their anxiety when encountering low-fluency processing of challenging materials, particu-

larly under time pressure (Caviola et al., 2017). However, in the absence of time constraints, students might derive greater benefits from engaging with more difficult items. It remains an open question whether the advantages of monitoring-based rewards would persist under such conditions.

11. Educational Implications

In real educational settings, educators often focus on praising students for their excellent performance. However, students who demonstrate strong metacognitive monitoring—i.e., those who can accurately assess their own performance—are frequently overlooked. The present study suggests that, compared to solely rewarding performance, acknowledging and encouraging metacognitive monitoring can offer significant benefits. In everyday learning contexts without external rewards, placing equal importance on metacognitive monitoring alongside performance could be a valuable strategy. This approach may help reduce performance anxiety, encourage more thorough out-of-class review, and foster a growth mindset. By emphasizing the role of metacognitive skills, educators can promote not only better performance but also a more resilient and self-aware approach to learning.

12. Conclusion Remarks

Providing students with monetary rewards is often viewed with caution due to its potential negative effects, such as performance-related anxiety and diminished intrinsic motivation. However, the present study demonstrates that it is not the monetary reward itself that leads to these issues, but rather the design of the reward scheme. Specifically, rewarding the alignment between memory performance and metacognitive accuracy resulted in significant memory enhancement and avoided the common pitfalls associated with traditional reward-based learning paradigms. Overall, this study suggests that for educators, metacognitive accuracy should be valued equally alongside pure academic performance. By focusing on how well students can assess their own learning and incorporating reward schemes that reinforce this skill, educators can foster both improved learning outcomes and a healthier learning environment.

Author Contributions: Conceptualization, S.L., C.K. and J.B.; methodology, S.L., C.K. and J.B.; software, S.L.; validation, S.L.; formal analysis, S.L.; investigation, S.L.; resources, S.L.; data curation, S.L.; writing—original draft preparation, S.L.; writing—review and editing, S.L., C.K. and J.B.; visualization, S.L.; supervision, C.K. and J.B.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the China Postdoctoral Science Foundation (2023M730303).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Faculty of Psychology at Beijing Normal University (file number 202201040001).(Approval date: 4 January 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: All the data are available from the OSF link: https://osf.io/3czqr/?view_only=e7d6ea6f9e174b06b794d660e4b1814e (accessed on 16 May 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

Ashcraft, M. H., & Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review*, *14*, 243–248.

Bandura, A., & Cervone, D. (1983). Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of Personality and Social Psychology*, *45*, 1017. [CrossRef]

Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, *46*, 610. [CrossRef]

Benware, C. A., & Deci, E. L. (1984). Quality of learning with an active versus passive motivational set. *American Educational Research Journal*, *21*, 755–765. [CrossRef]

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444. [CrossRef] [PubMed]

Caviola, S., Carey, E., Mammarella, I. C., & Szucs, D. (2017). Stress, time pressure, strategy selection and math anxiety in mathematics: A review of the literature. *Frontiers in psychology*, *8*, 1488. [CrossRef] [PubMed]

DeCaro, M. S., Thomas, R. D., Albert, N. B., & Beilock, S. L. (2011). Choking under pressure: Multiple routes to skill failure. *Journal of Experimental Psychology: General*, *140*, 390. [CrossRef]

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*, 627. [CrossRef]

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, *71*, 1–27. [CrossRef]

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist*, *26*, 325–346. [CrossRef]

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, *41*, 1040. [CrossRef]

Ellis, H. C., & Ashbrook, P. W. (1989). The "stat" of mood and memory research: A selective review. *Journal of Social Behavior and Personality*, *4*(2), 1.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175–191. [CrossRef] [PubMed]

Fröber, K., & Dreisbach, G. (2016). How performance (non-)contingent reward modulates cognitive control. *Acta Psychologica*, *168*, 65–77. [CrossRef]

Fröber, K., & Dreisbach, G. (2020). How Sequentially Changing Reward Prospect Modulates Meta-control: Increasing Reward Prospect Promotes Cognitive Flexibility. *Cognitive, Affective and Behavioral Neuroscience*, *21*, 534–548. [CrossRef]

Halamish, V., & Undorf, M. (2022). Why do judgments of learning modify memory? Evidence from identical pairs and relatedness judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *49*, 547. [CrossRef] [PubMed]

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology*, *92*, 316. [CrossRef]

Janes, J. L., Rivers, M. L., & Dunlosky, J. (2018). The influence of making judgments of learning on memory performance: Positive, negative, or both? *Psychonomic Bulletin & Review 25*, 2356–2364.

Knowlton, B. J., & Castel, A. D. (2022). Memory and reward-based learning: A value-directed remembering perspective. *Annual Review of Psychology*, *73*, 25–52. [CrossRef] [PubMed]

Kool, W., McGuire, J. T., Wang, G. J., & Botvinick, M. M. (2013). Neural and behavioral evidence for an intrinsic cost of self-control. *PLoS One*, *8*(8), e72626. [CrossRef] [PubMed]

Kuhbandner, C., Aslan, A., Emmerdinger, K., & Murayama, K. (2016). Providing extrinsic reward for test performance undermines long-term memory acquisition. *Frontiers in Psychology*, *7*, 79. [CrossRef]

Lachman, M. E., & Weaver, S. L. (1998). The sense of control as a moderator of social class differences in health and well-being. *Journal of Personality and Social Psychology*, *74*, 763. [CrossRef] [PubMed]

Lepper, M. R. (1973). Dissonance, self-perception, and honesty in children. *Journal of Personality and Social Psychology*, *25*(1), 65. [CrossRef]

Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Journal of Educational Psychology*, *97*, 184. [CrossRef]

Li, B., Hu, X., Shanks, D. R., Su, N., Zhao, W., Meng, L., Lei, W., Luo, L., & Yang, C. (2024). Confidence ratings increase response thresholds in decision making. *Psychonomic Bulletin & Review*, *31*(3), 1093–1102.

Li, B., Zhao, W., Zheng, J., Hu, X., Su, N., Fan, T., Yin, Y., Liu, M., Yang, C., & Luo, L. (2022). Soliciting judgments of forgetting reactively enhances memory as well as making judgments of learning: Empirical and meta-analytic tests. *Memory & Cognition*, 1–17.

Lin, H., Westbrook, A., Fan, F., & Inzlicht, M. (2024). An experimental manipulation of the value of effort. *Nature Human Behaviour*, *8*, 1–13. [CrossRef] [PubMed]

Liu, S., Zheng, Z., Kent, C., & Briscoe, J. (2022). Progressive retrieval practice leads to greater memory for image-word pairs than standard retrieval practice. *Memory*, *30*, 796–805. [CrossRef]

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174–179.

Miendlarzewska, E. A., Bavelier, D., & Schwartz, S. (2016). Influence of reward motivation on human declarative memory. *Neuroscience & Biobehavioral Reviews*, *61*, 156–176.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947. [CrossRef] [PubMed]

Murayama, K., & Kuhbandner, C. (2011). Money enhances memory consolidation–But only for boring material. *Cognition*, *119*, 120–124. [CrossRef] [PubMed]

Murayama, K., Matsumoto, M., Izuma, K., & Matsumoto, K. (2010). Neural basis of the undermining effect of monetary reward on intrinsic motivation. *Proceedings of the National Academy of Sciences*, *107*, 20911–20916. [CrossRef] [PubMed]

Musslick, S., Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2015, June 7–10). *A computational model of control allocation based on the expected value of control*. Reinforcement Learning and Decision Making Conference (Vol. 2015), Edmonton, AB, Canada.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13. [CrossRef] [PubMed]

Rivers, M. L., Janes, J. L., & Dunlosky, J. (2021). Investigating memory reactivity with a within-participant manipulation of judgments of learning: Support for the cue-strengthening hypothesis. *Memory*, *29*, 1342–1353. [CrossRef]

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, *40*(1), 99–124. [CrossRef]

Simons, J., Dewitte, S., & Lens, W. (2004). The role of different types of instrumentality in motivation, study strategies, and performance: Know why you learn, so you'll know what you learn! *British Journal of Educational Psychology*, *74*(3), 343–360. [CrossRef] [PubMed]

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 553. [PubMed]

Tauber, S. K., & Witherby, A. E. (2019). Do judgments of learning modify older adults' actual learning? *Psychology and Aging*, *34*, 836. [CrossRef]

Tekin, E., & Roediger, H. L., III. (2020). Reactivity of judgments of learning in a levels-of-processing paradigm. *Zeitschrift für Psychologie*, *228*(4), 278–290. [CrossRef]

Toppino, T. C., & Cohen, M. S. (2010). Metacognitive control and spaced practice: Clarifying what people do and why. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1480. [CrossRef]

Umemoto, A., & Holroyd, C. B. (2015). Task-specific effects of reward on task switching. *Psychological research*, *79*, 698–707. [CrossRef]

Van Loon, M. H., & Oeri, N. S. (2023). Examining on-task regulation in school children: Interrelations between monitoring, regulation, and task performance. *Journal of Educational Psychology*, *115*(3), 446. [CrossRef]

Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., & Deci, E. L. (2004). Motivating learning, performance, and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality and Social Psychology*, *87*, 246. [CrossRef]

Westbrook, A., Kester, D., & Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PLoS ONE*, *8*(7), e68210. [CrossRef] [PubMed]

Zhao, W., Li, B., Shanks, D. R., Zhao, W., Zheng, J., Hu, X., Su, N., Fan, T., Yin, Y., Luo, L., & Yang, C. (2022). When judging what you know changes what you really know: Soliciting metamemory judgments reactively enhances children's learning. *Child Development*, *93*, 405–417. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Listen or Read? The Impact of Proficiency and Visual Complexity on Learners' Reliance on Captions

Yan Li

School of Foreign Languages and Cultures, Jilin University, Changchun 130012, China; y_li20@mails.jlu.edu.cn

Abstract

This study investigates how Chinese EFL (English as a foreign language) learners of low- and high-proficiency levels allocate attention between captions and audio while watching videos, and how visual complexity (single- vs. multi-speaker content) influences caption reliance. The study employed a novel paused transcription method to assess real-time processing. A total of 64 participants (31 low-proficiency [A1–A2] and 33 high-proficiency [C1–C2] learners) viewed single- and multi-speaker videos with English captions. Misleading captions were inserted to objectively measure reliance on captions versus audio. Results revealed significant proficiency effects: Low-proficiency learners prioritized captions (reading scores > listening, $Z = -4.55$, $p < 0.001$, $r = 0.82$), while high-proficiency learners focused on audio (listening > reading, $Z = -5.12$, $p < 0.001$, $r = 0.89$). Multi-speaker videos amplified caption reliance for low-proficiency learners ($r = 0.75$) and moderately increased reliance for high-proficiency learners ($r = 0.52$). These findings demonstrate that low-proficiency learners rely overwhelmingly on captions during video viewing, while high-proficiency learners integrate multimodal inputs. Notably, increased visual complexity amplifies caption reliance across proficiency levels. Implications are twofold: Pedagogically, educators could design tiered caption removal protocols as skills improve while incorporating adjustable caption opacity tools. Technologically, future research could focus on developing dynamic captioning systems leveraging eye-tracking and AI to adapt to real-time proficiency, optimizing learning experiences. Additionally, video complexity should be calibrated to learners' proficiency levels.

Keywords: captioned videos; proficiency levels; language processing; desirable difficulties; video complexity; listening skills

1. Introduction

Videos have become a common and major way to provide input to learners in numerous second language (L2) classrooms for several decades (Chai & Erlam, 2008; Danan, 2004; Vanderplank, 2016). Many teachers use captioned videos to help improve the learners' listening abilities, as they expose learners to authentic oral language input that closely resembles real-life communication. However, there is a view that learners may simply read the captions and ignore the soundtrack, thus improving reading rather than listening skills (Vandergrift, 2004; Yeldham, 2024). Prior studies yield mixed findings: Low-proficiency learners consistently depend on captions as a crutch (Caimi, 2006; Vanderplank, 2019), while higher-proficiency learners integrate audio, visuals, and text (Taylor, 2005; Winke et al., 2013). However, contradictory evidence suggests even advanced learners may default to caption reliance (Chai & Erlam, 2008) if the task complexity exceeds their cognitive capacity.

Given the inconclusive and occasionally conflicting outcomes, further research is needed to investigate whether the learners mainly read the captions or actually listen to the videos. For one thing, understanding this dynamic is essential for optimizing the use of captioned videos in language teaching, as it enables educators to provide learners with appropriate video materials and effective instruction. For another, English educators need an efficient, user-friendly, and scientifically validated tool to assess students' reliance on captions. Such assessment could inform decisions about when to implement learner-appropriate difficulties, like gradually reducing caption support as proficiency increases. This investigation aligns with cognitive psychology's concept of "desirable difficulties" (Bjork & Bjork, 2011)—the principle that introducing optimal challenges (e.g., strategic caption removal) can enhance long-term learning.

In this study, I focused on whether learners primarily listened to the audio or read the captions while watching the videos. To accomplish this, paused transcription was used to investigate how much the participants relied on captions. The findings showed that low-proficiency learners predominantly relied on captions as a scaffold for comprehension, while high-proficiency learners demonstrated a greater ability to integrate auditory and textual cues. Additionally, video complexity, such as the presence of multiple speakers, can increase reliance on captions, particularly for low-proficiency learners. The outcomes aim to provide actionable thresholds for implementing desirable difficulties in caption scaffolding, ultimately informing the development of tailored instructional strategies that account for learners' proficiency trajectories and multimodal processing capacities.

2. Literature Review

In understanding how language learners process captioned videos, two prominent cognitive theories provide valuable frameworks: the cognitive theory of multimedia learning (Mayer, 2005) and the cognitive load theory (Sweller, 1988, 1994). These theories offer insights into how learners utilize different cognitive resources while watching videos with captions, such as processing auditory, visual, and textual information.

2.1. Theoretical Framework

2.1.1. Cognitive Theory of Multimedia Learning (CTML)

The cognitive theory of multimedia learning, developed by Mayer (2005), posits that people learn more effectively from words and pictures than from words alone. According to CTML, learning occurs through two distinct channels in the brain: the auditory channel (for processing spoken words) and the visual channel (for processing pictures and text). Mayer's theory suggests that effective learning happens when information from these two channels is presented in a way that reduces cognitive overload and facilitates the integration of the material into long-term memory.

The cognitive theory of multimedia learning entails three assumptions: the dual-channel assumption (i.e., information is processed through two channels: an auditory–verbal channel and a visual–pictorial channel), the limited capacity assumption (i.e., each channel has a limited capacity to process information), and the active processing assumption (i.e., humans actively engage in cognitive processing to construct mental representations of their experiences through attending to, organizing, and integrating incoming information).

The cognitive theory of multimedia learning emphasizes several principles demonstrated to affect the cognitive processing of information. One of them is the redundancy principle: including redundant information (e.g., presenting spoken text along with on-screen text) can lead to cognitive overload. In a study conducted by Mayer et al. (2001), native English-speaking participants watched a captioned animation explaining lightning while listening to concurrent narration in English. It turned out that they retained less

information than those who watched and listened to the same but uncaptioned animation. However, Sydorenko (2010) claimed that the principles above are "not borne out for second language learners" (p. 52). Foreign language learners might find that captions support their access to multimedia messages. The question is how language learners are different from native speakers in attending to the three channels of sound, image, and captions.

2.1.2. Cognitive Load Theory (CLT)

Developed by Sweller (1988, 1994), cognitive load theory (CLT) is a psychological theory that focuses on how the human cognitive system processes information. The theory posits that there are limits to the amount of information individuals can process effectively in their working memory.

Cognitive load refers to the mental effort and resources required to process information. There are three types of cognitive load: (1) Intrinsic cognitive load: This is the inherent complexity of the content being learned. Some subjects or concepts are naturally more complex and thus impose a higher intrinsic cognitive load. (2) Extraneous cognitive load: This is the cognitive load imposed by the instructional design, presentation, or learning environment. Poorly designed materials can lead to extraneous cognitive load, distracting learners from the main content and impeding learning. (3) Germane cognitive load: This type of cognitive load is directed toward building mental connections between new information and existing knowledge.

Working memory is one of the key concepts on which CLT is built. It is the part of the cognitive system responsible for temporarily holding and manipulating information. It can only process a limited amount of information at a time. When too many messages are presented, selection, organization, and integration occur.

The goal of CLT is to optimize learning by managing cognitive load. In multimedia presentations, teachers can find and expose learners to materials that consider the inherent complexity of subjects, the sequencing of topics, and the use of appropriate instructional strategies to manage cognitive load, ensuring that the materials do not overwhelm learners with excessive cognitive demands.

2.2. Empirical Evidence About How Learners Process Captioned Videos

The existing literature on captions focuses mainly on their benefits for second language learning, particularly in two main areas: general comprehension and vocabulary acquisition. Some studies have examined how captioned videos improve learners' listening comprehension skills. However, only nine studies focus on how learners process captioned videos.

2.2.1. Overview of the Captioned Video Processing

Of the nine studies outlined in the Supplementary Material, six focused on English learning. Three studies investigating foreign language learning (Russian and Spanish) by native English speakers were also included, as they provided valuable insights and inspiration for this research. Among these, six were cross-sectional, as they collected data on viewing behaviors at a single point in time (Cohen et al., 2002). The remaining three (Vanderplank, 1988, 2016, 2019) were longitudinal, as they collected data repeatedly from the same participants over periods of 9 weeks, 5–6 weeks, and 12 weeks, respectively.

Of the six cross-sectional studies, three included low-proficiency learners (Caimi, 2006; Sydorenko, 2010; Winke et al., 2013), two were composed of mixed lower- and higher-proficiency learners (Pujolà, 2002; Taylor, 2005), and the sixth one targeted upper-intermediate learners. The learners involved in those three longitudinal studies were all at an upper-intermediate to advanced level. Participants in Chai and Erlam's (2008) study were 10 Chinese ESL learners. One Cantonese-speaking student participated in

Sydorenko's (2010) study. Participants in other studies were mainly European foreign language speakers.

As to the videos used in the nine studies investigated here, seven used videos of various types and lengths, such as comedy series, movies, TV shows, and documentary clips. The exception was Taylor (2005), who used a segment from the videotape included within a textbook, while Caimi (2006) did not provide details about the specific kinds of videos that were utilized. While selecting captioned videos, the researchers of most studies, excluding Caimi, judged the selected videos suitable for the level of the students in their studies through various methods, such as vocabulary tests, instructors' evaluation, standard tests, etc.

Concerning the research techniques used to gather the data (see the Supplementary Material, column 4), most studies used introspective methods to assess participants' perceived attention distribution, usually immediately after the viewing task. These techniques included interviews, questionnaires, verbal reports, learner written reflective protocols, and learner diaries (hard-copy or online). Winke et al. (2013) also used eye tracking, via a sensor-based device which can tap into the learners' real-time processing of the videos and objectively record participants' attention distribution between captions and the visuals.

2.2.2. Main Results from the Cross-Sectional Studies

Regarding the findings of the cross-sectional studies, two main results were observed: First, low-proficiency learners mainly read captions, using captions as a scaffold to help them understand the videos; second, high-proficiency learners generally used visual, aural, and captioned cues simultaneously.

Caimi's (2006) study, which investigated how 15 Italian lower-intermediate EFL learners responded to captioned videos through interviews, found that students focused more on reading the captions than on listening. The study revealed that low-proficiency students became overwhelmed by the combination of visual, aural, and textual cues, leading them to rely heavily on reading the captions. The simultaneous presentation of visual, aural, and textual information impeded comprehension. This echoed the findings of Sydorenko (2010), who examined 26 second-semester US university Russian learners (high-level beginners) through questionnaires. The questionnaire results indicated that learners paid the most attention to captions, followed by the video and then the audio. Some participants had difficulty in attending to all three modalities. In the third study, Winke et al. (2013) used eye tracking and interviews to determine the total amount of time English-speaking learners of Chinese, Arabic, Russian, and Spanish spent on two documentaries dubbed and captioned in the target language. The results showed that these low-intermediate learners primarily focused on the captions, spending an average of 68% of their time fixating on the caption area when captions were present. Spanish and Russian learners exhibited different caption-reading behaviors compared to Arabic and Chinese learners, with fixation times of 63% and 67% for Spanish and Russian learners, respectively, and 75% and 68% for Arabic and Chinese learners. Notably, Chinese learners spent 62% of their time looking at captions when the video content was familiar, compared to 74% when the content was unfamiliar. The Chinese learners reported spending most of their time reading the captions because the Chinese characters were difficult to process. Winke et al. (2013) also mentioned that some native language studies investigated how readers read still pictures with captions (e.g., Kennedy et al., 2000; Rayner et al., 2001; and Underwood et al., 2004). The overall picture emerging from these studies has shown that readers tend to look briefly at the picture first (as few as three eye fixations), the caption next, and then the picture again.

In Pujolà's (2002) multi-proficiency study, which gathered data through observation, verbal report, and interview, the less-proficient learners also relied more on the

captions. Two students even mentioned "as if they were doing a reading task instead of listening" (Pujolà, 2002, p. 253). However, the more proficient learners used captions less often because they were able to follow the sound. They only turned to the captions when they experienced comprehension difficulties. In the other multi-proficiency study, Taylor (2005) examined 35 US Spanish learners of intermediate and upper-intermediate level. In their reflective written protocols, some first-year students reported that they found the captions distracting and had difficulty using image, sound, and captions simultaneously. Experienced learners reported they could use three channels together while watching captioned videos.

These five studies indicate that the way foreign language (FL) learners process captioned videos is closely related to their proficiency levels. As we mentioned above, the two main trends that emerged are that less proficient learners tend to mainly read the captions and that more proficient learners generally typically use visual, aural, and captioned cues flexibly. In a recent study, Pannatier and Bétrancourt (2024) investigated the impacts of subtitle language and learners' foreign language (FL) proficiency on learning from subtitled academic videos. 131 French-speaking students with different English-proficiency levels watched a video under three conditions: English same- language subtitles (SLS), French native language subtitles (NLS), or no subtitles. Using a between-group design, they measured comprehension (retention, inference, and transfer), self-reported cognitive load, and situational interest. Results show no subtitle effects but a significant proficiency main effect: advanced learners outperformed lower-proficiency peers on inference/transfer tasks and reported reduced extraneous cognitive load. This aligns with cognitive load theory, suggesting that higher proficiency mitigates linguistic processing demands, freeing resources for integrating audio, visual, and textual inputs. These findings resonate with the trends above: advanced learners' lower cognitive load supports multi-modal integration, whereas elementary learners, facing higher load, may prioritize captions as a compensatory strategy. However, an exception was Chai and Erlam's (2008) study, which involved 10 native Chinese speakers with upper-intermediate English proficiency. This finding contradicted the second trend that more proficient learners used multiple channels. Eight participants reported that they "paid most of their attention to the captions" (Chai & Erlam, 2008, p. 34). The results suggested that teachers should avoid using captions when incorporating videos to enhance L2 learners' listening skills. Unlike other cross-sectional studies, Chai and Erlam (2008) divided 20 participants into a no-captions group and a captions group. Their procedure also differed from other studies. Participants first watched the video in their native language (Chinese) as a scaffold to help them understand the storyline, a method that is uncommon in both research and classroom contexts. After viewing, participants were shown 21 words and phrases from the video script and instructed to "try and listen for these words and the sentences in which they occurred" (Chai & Erlam, 2008, p. 28). The researchers explained that their study focused on individual words and phrases rather than overall comprehension. This might have subtly encouraged participants to focus more on the captions.

2.2.3. Main Results from the Longitudinal Studies

All three longitudinal studies were conducted by Vanderplank (1988, 2016, 2019). Their results were consistent and demonstrated a trend of continuous improvement. All three studies involved proficient learners at the upper-intermediate to advanced level. The results can be categorized into three groups: (1) relatively less proficient students among the high-level learners used captions as a scaffold; (2) proficient learners alternated between sound, image, and captions, using captions to reinforce their understanding; and (3) most learners exhibited changes in their viewing behaviors over time.

In Vanderplank's (1988) study, fifteen upper-intermediate to advanced European ESL learners watched authentic TV programs of various genres for one hour per week over nine weeks. The findings indicated that some learners alternated between spoken and caption cues as needed, while others simultaneously utilized spoken, visual, and caption cues, similar to the behaviors of proficient learners in Pujolà's (2002) and Taylor's (2005) studies. Initially, some learners found the captions distracting, but this effect gradually diminished over time. Only two learners reported a continued need or desire to use captions after nine weeks. They felt that watching uncaptioned TV programs would have encouraged them to listen more attentively and improve their listening abilities. Additionally, some participants struggled with accents and the excessive amount of information, which hindered their engagement and comprehension. The findings also suggested that captioned programs "may be of limited value for low-level learners" (Vanderplank, 1988, p. 272). Eight low-intermediate to advanced Arabic students in the study group found that the captions changed too quickly.

In Vanderplank's (2016) study, 18 learners of European foreign languages at the upper-intermediate to advanced level selected from a wide range of films of various genres and watched them at their own pace and control. In this study, Vanderplank observed similar changes in viewing behaviors among the learners, which aligned closely with those identified in his earlier research (Vanderplank, 1988). All participants aimed to watch films confidently, and the presence of captions shifted their focus from "viewing for enjoyment" to "conscious learning". The participants reported using captions when needed. The more advanced participants used captions primarily as a backup when they struggled to understand the speakers. As the study progressed, they gradually reduced their reliance on captions. All participants noted in their diaries that many films would have been difficult to appreciate without captions. However, two lower-level participants reported still relying on captions by the end of the study. The final questionnaires in Vanderplank (2016) indicated that most learners experienced an overall increase in confidence by the end of the study. Many participants realized that they relied more on reading captions than on listening. As their confidence grew, they used captions flexibly for "reassurance".

In Vanderplank's (2019) study, thirty-six intermediate-advanced learners of French, German, Italian, and Spanish watched authentic movies of various genres. This longitudinal study consisted of two trials. The first trial was the second longitudinal study, involving 18 learners over six weeks, while the second trial, involving 18 participants, was extended to twelve weeks. The findings closely aligned with those of the previous two longitudinal studies and mirrored the outcomes observed in Pujolà (2002) and Taylor (2005). The participants tended to read captions at the start of viewing, particularly when encountering unfamiliar accents, rapid dialogue, or content beyond their proficiency level. Eleven participants expressed concerns about becoming too reliant on captions. Nevertheless, their viewing behaviors evolved over time to "maximize the value of the tri-modal input of sound, text, and visuals" (Vanderplank, 2019, p. 417). Initially, only four participants expressed negative views about the value of captions. However, their attitudes shifted from "hostile" to "positive" over time because, without captions, they would not have been able to follow the speech or understand the plots. Most participants noted that their confidence and ability to watch uncaptioned films improved as they watched more films. Additionally, the findings suggested that a recreational mindset "appeared to hinder the potential benefits of captioning in both formal and informal viewing contexts" (Vanderplank, 2019, p. 408).

2.3. Summary of Previous Research

Thus far, prior studies have investigated how language learners process captioned videos both in and beyond classroom settings using cross-sectional or longitudinal data. One of the two main findings is that low-proficiency learners mainly read captions, using them as a scaffold to help them understand the videos. This was observed in three cross-sectional studies involving only low-proficiency learners (Caimi, 2006; Sydorenko, 2010; Winke et al., 2013). This was also found among the less proficient learners in Pujolà's (2002) multi-proficiency study. In the other multi-proficiency study, Taylor (2005) found that more first-year students than third-year students struggled to attend to sound, image, and captions simultaneously, though it was unclear how they resolved this issue.

The second main result, emerging from Pujolà's (2002) and Taylor's (2005) studies, is that higher-proficiency learners flexibly attended to three channels: sound, image, and captions. Learners in Pujolà's (2002) study used captions to reinforce their comprehension when needed. However, Chai and Erlam's (2008) study, which involved 10 native Chinese speakers with proficient English, contradicted the second result, as more proficient learners did not consistently use multiple channels. Eight participants reported that they "paid most attention to the captions" (Chai & Erlam, 2008, p. 34).

The results of the three longitudinal studies (Vanderplank, 1988, 2016, 2019) can be categorized into three groups. First, less-proficient learners relied more heavily on captions as a scaffold, a finding consistent with Chai and Erlam's (2008) study. Second, proficient learners alternated between sound, image, and captions, using captions to reinforce their understanding, echoing the findings of Pujolà (2002) and Taylor (2005). A novel insight from Vanderplank (2019) advanced the research by showing that most learners' viewing behaviors evolved over time. They tended to read captions at the start of viewing. At this stage, the participants' viewing behaviors resembled those observed in Chai and Erlam's (2008) study.

In summary, this review demonstrates that previous studies do not provide a definitive conclusion about how low-proficiency and, particularly, high-proficiency learners process captioned videos. Given the inconclusive and occasionally conflicting findings, further research is needed to draw definitive conclusions.

2.4. Research Gaps

2.4.1. Research Techniques

The tools utilized in most of the nine studies were introspective, mainly questionnaires (e.g., Sydorenko, 2010 and Vanderplank, 2019), verbal reports (e.g., Pujolà, 2002), or interviews (e.g., Chai & Erlam, 2008; Taylor, 2005; Sydorenko, 2010; and Vanderplank, 1988). Despite their strengths, they are not capable of capturing the temporal fluctuations in cognitive processes. Their retrospective nature of reporting "can reduce their reliability" (Yeldham, 2017, p. 18). Winke et al. (2013) used eye-tracking technology to objectively assess participants' real-time attention distribution between captions and the videos. However, eye-tracking can be used to investigate how learners attend to videos and captions, but not audio. There is a need for measurements that directly indicate cognitive processing during the viewing of captioned videos. Therefore, I need multiple techniques (such as eye-tracking, paused transcription, and interviews) to complement the deficiencies of each technique and to triangulate the findings from each one. The overall picture can give us a more accurate and objective description regarding how learners pay attention to sound, image, and captions when watching captioned videos.

2.4.2. Basis for Selecting Videos

In some studies, the basis for the selection of videos was not specific, and criteria were too random. However, whether the videos' difficulty level matches participants' proficiency is a major limitation to researching how second or foreign language learners process captioned videos. Undoubtedly, if the content difficulty level is too far above participants' proficiency, they tend to read the captions more, while according to Krashen's (1982) comprehensible input theory, learners should be provided with input a bit beyond their current level of competence. Therefore, the selected videos should not be so challenging that learners are driven to read the captions. Additionally, other factors which should be taken into consideration include topic familiarity, delivery speed, speaker clarity, accent familiarity, lexical–syntactic difficulty, the correspondence between the visual, spoken, and caption cues, and especially, the level and type of visual movements (talking-head or multi-speaker video) (Yeldham, 2018b). One or two voices in a talking-head video do not pose a major challenge to both groups, but in videos "where there are three or more voices, the demands upon the listener are inevitably stepped up" (Field, 2019, p. 293).

2.4.3. Targeted EFL Learner Groups

The existing research primarily focused on full-caption videos' viewing behaviors of learners whose mother tongues were mostly Indo-European languages. Thus, further research is needed to know the viewing behaviors of learners in an EFL context in which there exist distinct L1–L2 orthographic, phonological, and semantic differences, such as native Mandarin-speaking Chinese EFL learners. As noted by Winke et al. (2013), "it is, therefore, worthwhile to investigate whether languages with different writing systems, phonological features and levels of cognitive demand differentially affect language learners' use of captions" (p. 257).

2.5. Research Questions

To address the gaps identified above, the present study focuses on the following two research questions.

1. Do Chinese students of high and low proficiency primarily focus on listening to the audio or reading captions when watching captioned videos?
2. Do they rely more on captions due to increased visual complexity in multi-speaker videos compared to single-speaker videos?

3. Methodology

3.1. Participants

The study involved 64 Chinese students from a university in Dalian, Liaoning Province, China. The participants were students from non-English disciplines, such as software engineering, bioengineering, communications, and economics. They had learned English in traditional classrooms, where the focus was mainly on vocabulary, grammar, and reading, with the aim of improving test performance rather than communicative skills.

While a formal power analysis was not conducted due to the exploratory nature of this study examining attention allocation patterns, the sample size was determined by adhering to the widely cited guideline from Larson-Hall (2015), which recommends a minimum of 30 participants per group for detecting medium-sized effects in between-group comparisons. After excluding participants with incomplete or low-quality data, the final sample included 31 low-proficiency (A1/A2 CEFR) and 33 high-proficiency (C1/C2 CEFR) learners.

The low-proficiency group consisted of second-year students (mean age = 19.9, SD = 0.3), while the high-proficiency group included third- and fourth-year students (mean age = 20.3, SD = 0.47). All participants indicated that they had normal or corrected

vision and reported no hearing impairments. The experiment was conducted with the written consent of all participants, who each received CNY 100 (approximately USD 14) as compensation.

3.2. Materials

Each group (low proficiency and high proficiency) watched two videos appropriate to their proficiency level: one single-speaker video and one multi-speaker video. The choice of these video types, as opposed to selecting documentaries as Winke et al. (2013) did, was based on their closer resemblance to real-life communication scenarios.

The low-proficiency group watched an 18 min single-speaker video on "Traffic, Weather, and Pets", while the multi-speaker video for this group focused on job hunting. The high-proficiency group watched a 15 min single-speaker video on "Teenagers, Adulthood, and Procrastination", while the multi-speaker video was an excerpt from a US sitcom.

The videos were selected based on criteria suggested by Yeldham (2018b), including topic familiarity, delivery speed, speaker clarity, accent familiarity, lexical–syntactic difficulty, and the correspondence between visual, spoken, and caption cues, with particular attention to the level and type of visual movements. Two instructors independently assessed the videos for their appropriateness in terms of difficulty and agreed that the videos were well-matched to the proficiency levels of the students. Despite inherent challenges in video selection, this study systematically controlled key variables (e.g., linguistic complexity) to minimize confounding factors, thereby strengthening the methodological rigor of the design.

3.3. Instrument: Paused Transcription

In paused transcription, a long recording is paused at pre-arranged irregular intervals. Each pause is followed by a black screen that lasts 10–15 s, during which learners write down the final 4–5 words of the sentence they have just heard. This approach can tap into the learners' decoding processes and allows researchers to obtain more immediate feedback on how L2 listening is processed (Patterson, 2019). This is based on psycholinguistic evidence from Jarvella (1971), which suggests that listeners can briefly retain a verbatim record of the words they hear until the onset of the following clause (Field, 2008, p. 419). This technique was used by Field (2008) and Yeldham (2016) to investigate L2 English learners' decoding of content words and function words. Additionally, Yeldham (2018a) employed this method to examine the influence of formulaic language on L2 listener decoding. The validity of the paused transcription technique was analyzed and confirmed by Patterson (2019), who demonstrated its effectiveness in capturing L2 learners' real-time decoding processes.

In this study, the paused transcription process largely followed the methods used by Field (2008) and Yeldham (2016, 2018a). Each captioned video was paused at random natural breaks, and participants transcribed the last 4–5 words of the sentence they had just heard. Each video included 20 random pauses. Of these, 10 captions matched the audio, while intentional errors were introduced into the remaining 10 captions. The underlying assumption was that individuals who relied on captions would primarily consider the content they had read as their reference. Consequently, when encountering misleading captions, they would transcribe what appeared on the screen. In contrast, those who could accurately transcribe despite the mismatch between on-screen captions and auditory input could be considered less dependent on captions.

As mentioned above, participants only needed to transcribe the last 4–5 words of the sentence they had just heard. This was also the reason why I did not explicitly test

their working memory. According to Miller's (1956) "The Magical Number Seven, Plus or Minus Two", the typical capacity of working memory for adults is around 7 ± 2 units of information. Given this well-established concept, I assumed that asking participants to recall only a few words from the end of a sentence would not exceed their typical working memory capacity. Therefore, this task was designed more to assess language processing rather than directly measuring working memory performance.

The transcription sections were carefully selected and edited based on several considerations: (1) Pauses were arranged irregularly at natural breaks, primarily at sentence endings, to prevent predictability and maintain video continuity. (2) The duration was set according to participants' proficiency: 15 s for the low-proficiency group and 10 s for the high-proficiency group, allowing for the transcription of approximately six words. (3) Each pause was inserted after a segment of at least 10 s of continuous playtime to avoid affecting viewing and understanding. (4) Participants only needed to write down the last 4–5 words to avoid extra cognitive load. (5) Edited captions (or misleading captions) were subtle and still made sense both grammatically and contextually without hindering understanding. (6) A black screen followed each misleading caption during transcription, and a countdown reminder played three seconds before the screen disappeared to signal video resumption.

There are four strategies for modifying captions: adding extra words, dropping words, replacing words, and misplacing words. In this study, the misleading captions were mainly created by replacing words, adding extra words, and deleting words.

(1) Adding words	
Original Content	Misleading captions
I think everyone should try to walk more.	I think everyone should try to walk out more.
I don't have time right now.	I don't have much time right now.
Taking control of your life	Taking control of your own life
(2) Deleting words	
Original Content	Misleading captions
She lived with us for over 10 years.	She lived with us for 10 years.
you need to tell them the truth	you need to tell the truth
This is one technique that you can employ in your own studies.	This is one technique that you can employ in your studies.
(3) Replacing words	
Original Content	Misleading captions
the subway stations are dirty and a little scary.	the subway stations are dirty and a little scaring.
They were smart and fun.	They were smart and funny.
You have to take turns.	You ought to take turns.

To determine whether learners relied more on listening to the spoken audio or reading the captions, I counted students' written transcriptions of video segments containing edited captions. A transcription aligned with the edited captions or the audio awarded 1 point. For example, if students accurately transcribed "They were smart and fun" (matching the original correct version from the audio) instead of the misleading captions "They were smart and funny", they received 1 point for listening; if students wrote "They were smart and funny" (matching the misleading captions) instead of the original correct version from the audio "They were smart and fun", they received 1 point for reading. Minor spelling mistakes were disregarded. For instance, if students transcribed the edited caption "They were smart and funny" as "They were smart and funy" (missing one letter 'n'), they would still receive 1 point for reading because the key criterion here was whether they included the "y" from the edited caption that was not present in the audio, and spelling errors did

not affect scoring in this context. However, if learners' transcriptions matched neither the edited captions nor the original audio, they would be marked as "wrong". For example, if the speaker in the video said "I don't have time now" while the edited caption displayed "I don't have much time now", but learners wrote "I don't have any time now", this would be marked as "wrong". However, such cases were relatively rare and had little impact on addressing the two research questions, so they were excluded.

3.4. Data Collection

The paused transcription test was conducted in a tiered lecture hall at a university. This choice was made for three reasons. First, one of the research objectives was to find an efficient, user-friendly, and scientifically sound method for teachers to assess students' reliance on captions. This would enable teachers to select suitable captioned videos and decide when or if to remove captions. Therefore, a tiered lecture hall, accessible to all teachers, was the best option. Second, this approach allowed for the simultaneous testing of all participants, increasing efficiency compared to individual testing. Third, this method was cost-effective because the lecture hall was equipped with a projector, a sound system, and a large screen.

Participants first filled out a background questionnaire, which included demographic information (name, gender, age, school, major, and grade), as well as details about their academic and language background, including the length of English study, their English score in the College Entrance Examination, and their scores on standardized tests (CET 4/CET 6). This was followed by an example and instructions explaining the order of the tasks. At the bottom of the first page was an informed consent form, which informed students that the data from the experiment would be used solely for scientific research. Students were required to sign it.

Before the experiment, participants were explicitly instructed to prioritize listening over captions. This was to encourage participants to listen without relying on captions, as they would in real-life interactive situations. They were instructed to write down only what they had heard.

Before the experiment officially began, an example video was shown to demonstrate the procedure and ensure that everyone understood the process. A paused transcription test was then administered. Participants first watched an English talking-head video with English captions. Each captioned video was paused at natural breaks. A black screen followed each pause, and participants transcribed the last 4–5 words they had just heard. Participants then watched the English multi-speaker videos. The process was the same as watching the single-speaker videos. Then, students were free to leave.

4. Results

Transcriptions were scored in accordance with the criteria outlined in the research instrument section. The quantitative data were analyzed using SPSS 29. The Shapiro–Wilk test was applied to assess whether the transcription scores (reading and listening) for each group (high/low proficiency) and video type (single/multi-speaker) followed a normal distribution. Based on that, either parametric tests (such as the *t*-test) or non-parametric tests (like the Wilcoxon signed-rank test) were employed to compare two sets of data and ascertain whether a significant difference existed between them. Moreover, the effect size was considered. Cohen's *r*, a frequently used effect-size metric, particularly suitable for non-parametric data, was utilized ($r = 0.10$ as a small effect, $r = 0.30$ as a medium effect, and $r = 0.50$ as a large effect).

4.1. Findings of Research Question 1

The results of the Shapiro–Wilk test revealed that, for the low-proficiency group, both reading scores ($W = 0.914$, $p = 0.016$) and listening scores ($W = 0.873$, $p = 0.001$) significantly deviated from normality ($p < 0.05$ for both). Similarly, the high-proficiency group showed non-normal distributions for reading ($W = 0.876$, $p = 0.002$) and listening ($W = 0.912$, $p = 0.013$). Given the non-normal distribution of the data, non-parametric tests were deemed appropriate for further analysis. To examine whether students of high and low proficiency relied more on reading captions or listening to audio during video viewing, a Wilcoxon signed-rank test was selected to compare the paired reading and listening scores within each group. Effect sizes were calculated using Cohen's r to quantify the magnitude of differences observed.

Transcription scores for listening and reading in single- and multi-speaker videos (total score: 20 points) were compared. Descriptive statistics for the low-proficiency group's transcription scores revealed a marked preference for reading captions over listening to audio. As shown in Table 1, the mean reading score ($M = 10.52$, $SD = 3.24$) was substantially higher than the mean listening score ($M = 5.81$, $SD = 2.94$). A Wilcoxon signed-rank test confirmed this difference was statistically significant ($Z = -4.55$, $p < 0.001$), with a large effect size (Cohen's $r = 0.82$), indicating that low-proficiency learners relied overwhelmingly on captions during video viewing, as evidenced by the fact that the median reading scores (11) were double the median listening scores (5), indicating their strong preference for captions.

Table 1. Listening vs. reading transcription scores for both groups and Wilcoxon signed-rank test results.

Group	Transcription Scores for Listening (*M*, *SD*)	Median	Transcription Scores for Reading (*M*, *SD*)	Median	*Z*	*p*	Cohen's *r*
Low (n = 31)	5.81 (2.94)	5	10.52 (3.24)	11	−4.55	<0.001	0.82
High (n = 33)	12.27 (3.12)	11	6.70 (3.84)	6	−5.12	<0.001	0.89

Conversely, high-proficiency participants demonstrated the opposite pattern. As presented in Table 1, their mean listening score ($M = 12.27$, $SD = 3.12$) out of 20 points significantly surpassed the mean reading score ($M = 6.70$, $SD = 3.84$). A Wilcoxon signed-rank test indicated a statistically significant difference between listening and reading scores ($Z = -5.12$, $p < 0.001$), with an equivalently large effect size (Cohen's $r = 0.89$). Additionally, the median listening score (11) was nearly double the median reading score (6), further indicating a strong preference for auditory input. These findings suggest that high-proficiency learners prioritized auditory input over captions during video viewing.

These divergent patterns across proficiency levels provide robust evidence that language proficiency moderates learners' reliance on captions versus audio. Low-proficiency students depended significantly on captions (median reading score = 11 and median listening score = 5; reading > listening, $Z = -4.55$, $p < 0.001$, $r = 0.82$), whereas high-proficiency students prioritized audio (median reading score = 6 and median listening score = 11; listening > reading, $Z = -5.12$, $p < 0.001$, $r = 0.89$). These findings align with theoretical frameworks positing that lower proficiency necessitates scaffolding via textual support, while higher proficiency enables integrated audio–visual processing.

4.2. Findings of Research Question 2

The second research question examines whether learners rely more on captions in multi-speaker videos than in single-speaker videos due to the increased visual complexity.

The normality of the reading and listening scores for both single-speaker and multi-speaker videos was assessed using the Shapiro–Wilk test. For the low-proficiency group, reading scores in single-speaker videos ($W = 0.901$, $p = 0.007$) and multi-speaker videos ($W = 0.892$, $p = 0.004$) significantly deviated from normality. Similarly, for the high-proficiency group, reading scores in both single-speaker ($W = 0.884$, $p = 0.003$) and multi-speaker videos ($W = 0.891$, $p = 0.005$) were non-normally distributed. These results ($p < 0.05$ for all tests) violated the assumptions of parametric tests, necessitating non-parametric analyses. For between-video comparisons, additional Wilcoxon signed-rank tests were conducted to evaluate caption reliance in multi-speaker versus single-speaker videos. Effect sizes were calculated using Cohen's r, interpreted as small ($r = 0.10$), medium ($r = 0.30$), or large ($r = 0.50$) effects based on established thresholds.

To address this question, I calculated and compared the reading scores between single-speaker and multi-speaker videos within each proficiency group. To be specific, two statistical tests were conducted: (a) For the low-proficiency group, the transcription scores for reading from both single-speaker and multi-speaker videos were calculated and compared to determine if there was a significant difference. (b) Similarly, for the high-proficiency group, the transcription scores for reading from the two types of videos were calculated and compared.

Table 2 revealed low-proficiency learners' higher reading scores in multi-speaker videos ($M = 6.06$, $SD = 1.93$) compared to single-speaker videos ($M = 4.94$, $SD = 1.84$). A Wilcoxon signed-rank test confirmed this increase was statistically significant ($Z = -4.21$, $p < 0.001$), with a large effect size ($r = 0.75$). This suggests low-proficiency learners relied more on captions when processing multi-speaker content. The median reading score increased from five (single-speaker) to six (multi-speaker), reflecting a consistent upward shift in typical caption reliance. Despite no change in the median for high-proficiency learners (three $\rightarrow$ three), high-proficiency learners, while still prioritizing listening overall, showed a small but significant increase ($p = 0.003$) in caption reliance during multi-speaker videos ($M = 3.70$, $SD = 1.89$) compared to single-speaker videos ($M = 3.09$, $SD = 1.97$), with a moderate practical effect ($r = 0.52$).

Table 2. Comparison of reading scores for both videos and Wilcoxon signed-rank test results.

Group	Single-Speaker (Reading)	Median	Multi-Speaker (Reading)	Median	*Z*	*p*	Cohen's *r*
Low (n = 31)	$M = 4.94$, $SD = 1.84$	5	$M = 6.06$, $SD = 1.93$	6	−4.21	<0.001	0.75
High (n = 33)	$M = 3.09$, $SD = 1.97$	3	$M = 3.70$, $SD = 1.89$	3	−2.97	0.003	0.52

This suggests that increased visual complexity (multi-speaker content) amplifies caption reliance across proficiency levels, but the impact is modulated by learners' language skills. Low-proficiency learners depend more heavily on captions while high-proficiency learners showed a small but significant increase in caption reliance while watching multi-speaker videos.

5. Discussion

The findings of this study align with, and in some cases extend, the existing literature on how learners of different proficiency levels process captioned videos, offering subtle insights into the interplay between proficiency levels, cognitive load, and instructional design.

5.1. Interpretation of Findings

Consistent with the cross-sectional studies (Caimi, 2006; Sydorenko, 2010; Winke et al., 2013), low-proficiency participants in this study demonstrated a strong reliance on captions with reading scores significantly surpassing listening scores ($Z = -4.55$, $p < 0.001$), which supports the notion that low-proficiency learners use captions as a scaffold to aid comprehension. The results echo Caimi's (2006) observation that low-proficiency learners often become overwhelmed by the tri-modal input, leading them to focus primarily on reading captions. Similarly, Sydorenko's (2010) finding that low-proficiency learners prioritize captions over audio and visual cues is reinforced by the current data.

In contrast, high-proficiency participants in this study demonstrated superior auditory processing (listening > reading, $Z = -5.12$, $p < 0.001$), reflecting their ability to allocate cognitive resources efficiently. This finding aligns with Pujolà's (2002) and Taylor's (2005) studies, which found that higher-proficiency learners are more adept at flexibly integrating visual, auditory, and textual cues. However, the slight increase in incongruent captions when watching multi-speaker videos suggests that even advanced learners may occasionally rely on captions due to increased cognitive load, such as when tracking multiple speakers. This nuanced finding adds to the literature by highlighting that proficiency level is not the sole determinant of reliance on captions; situational factors, such as video complexity, also play a role. The results also resonate with Vanderplank's (1988, 2016, 2019) longitudinal studies, which found that learners' reliance on captions evolves over time. While low-proficiency learners in this study consistently relied on captions, high-proficiency learners demonstrated a more balanced use of auditory and textual cues, similar to the proficient learners in Vanderplank's studies who used sound, image, and captions as needed. The current study's findings suggest that learners' processing strategies can vary based on video type, even within a single experimental session, echoing Vanderplank's observation that viewing behaviors evolve with exposure and experience.

However, the results contrast with Chai and Erlam's (2008) study, which found that even high-proficiency learners relied heavily on captions. This discrepancy may be attributed to differences in experimental design. Chai and Erlam's study focused on word and phrase recognition, which may have inadvertently encouraged participants to prioritize captions. In contrast, the current study employed a more naturalistic approach, enabling participants to process captioned videos holistically. This methodological difference underscores the importance of task design in shaping learners' reliance on captions.

Visual complexity, operationalized through multi-speaker videos, increased caption reliance among low-proficiency learners (reading scores increased by 23.7%, $Z = -3.85$, $p < 0.001$). This aligns with CLT's prediction that heightened intrinsic load (e.g., tracking multiple speakers) amplifies dependence on extraneous supports like captions. High-proficiency learners, however, exhibited only a modest increase in caption reliance (19.7%, $Z = -2.97$, $p = 0.003$), suggesting their superior ability to parse complex auditory–visual streams without over-relying on text. This divergence underscores proficiency as a critical moderator of cognitive load management.

5.2. Implications: Adaptive Strategies for Captions and Videos

The findings of this study suggest tailored strategies to adjust multimedia inputs for learners at different proficiency levels. These strategies aim to balance cognitive load while fostering skill progression, directly addressing the observed differences in caption reliance and auditory processing.

In practice, educators could design tiered caption removal protocols to align with proficiency thresholds and phase out captions dynamically. For example, low-proficiency learners can start with full captions, then highlight keywords (e.g., idioms and complex

grammar) while dimming others, and finally progress to no captions—a progression that nudges learners toward selective attention as proficiency improves. High-proficiency learners can be challenged with partially captioned or uncaptioned materials to foster listening skill development. Also, educators may consider calibrating video complexity by prioritizing single-speaker content for low-proficiency learners to help mitigate cognitive overload, while multi-speaker videos could be reserved for advanced learners as potential strategic listening challenges.

In addition, instead of a one-size-fits-all approach, teachers can use simple tools (like adjustable caption opacity in PowerPoint or YouTube) to decrease opacity from 100% to 30% and phase out captions in sync with student progress, forcing a strategic shift from text-dependent decoding to auditory-focused processing. For instance, the progress could start with full captions in Week 1 to build learner confidence, followed by reducing caption opacity to 50% while highlighting keywords in Week 3 to train selective attention. By Week 6, learners would transition to no captions for familiar topics, gradually fostering auditory independence through carefully calibrated challenges.

Finally, to address the challenge of multi-speaker scenes overwhelming beginners (as shown in the data), this study advocates for a complexity-aware design framework that tags videos with "complexity scores" (e.g., 1–5 stars) and pairs them with adaptive caption support. For instance, a five-star debate video, characterized by fast-paced verbal exchanges and diverse voices, would retain highlighted keywords to scaffold comprehension, while a two-star monologue with clear speech could be presented without captions to encourage auditory engagement. This approach leverages visual cues to dynamically balance cognitive load, allowing learners to progressively tackle more challenging content while maintaining focus on critical linguistic elements. By aligning multimedia complexity with tiered support systems, educators can systematically guide learners from reliance on textual aids to independent audio processing, fostering skill development through intentional difficulty calibration.

5.3. Limitations and Future Directions

While this study provides insights into how low- and high-proficiency learners allocate their attention and how caption types influence learner viewing behaviors, several limitations should be acknowledged alongside critical questions for future investigation.

First, I recognize that uncontrolled variables—including video content characteristics (e.g., humor) and learner factors (individual interest and prior knowledge)—may have mediated the observed effects. The video selection process prioritized linguistic appropriateness (grammar structures, lexical difficulty, and speech rate), matching participants' proficiency levels, followed by secondary controls for accent consistency, topic familiarity, and comparable duration. While this approach ensured pedagogical suitability, it necessarily constrained my ability to perfectly balance other potentially influential factors, particularly stylistic elements and nuanced topic differences between single-speaker and multi-speaker videos. Future research could implement stricter content-matching protocols and include measures of individual differences to better isolate the specific effects of caption variations. These refinements would help disentangle the observed effects from potential confounders while preserving the ecological validity of authentic video materials.

Second, the transcription-based measurement, while providing direct evidence of attention allocation between audio and captions, cannot capture more subtle cognitive processes like the depth of processing or comprehension. Although eye-tracking technology alone cannot distinguish whether learners are actively reading captions or listening to the audio, its incorporation could synergistically address the second research question. Future research could adopt an innovative methodology integrating eye-tracking technology with

a modified paused transcription task. Specifically, rather than verbatim text transcription, participants would engage in answering multiple-choice questions during pre-designated video pauses while maintaining head stability for eye-tracking calibration. This hybrid approach would enable researchers to systematically examine how learners dynamically allocate cognitive resources across multiple modalities (audio, visual, and captions) during multimedia engagement.

Third, the data in this study do not provide a definite answer to what difficulties are desirable in second/foreign language learning. Specifically, the observed preferences for modality-specific attention, whether toward captions or audio, reflect strategic behaviors rather than confirmed learning benefits. To address this gap, future research could explore this concept further. For instance, a follow-up study might examine how the development of listening abilities varies between low- and high-proficiency learners when exposed to captioned versus non-captioned materials. Such an investigation would help determine two critical issues: (1) whether removing captions or reducing caption support truly benefits high-proficiency learners, and (2) whether low-proficiency learners could likewise gain advantages from caption removal under specific conditions. Importantly, these inquiries align with the findings of Yue et al. (2013), who cautioned that learner preferences for fluent conditions (e.g., full captions) do not necessarily correlate with optimal learning outcomes. In their study, moderate discrepancies between narration and on-screen text in multimedia learning were found to create desirable difficulties, even though learners consistently preferred on-screen text to match the audio narration.

Ideally, future research on captions should not only analyze strategic differences in caption usage among learners at varying proficiency levels, but also focus on developing dynamic captioning systems that leverage eye-tracking and AI-driven proficiency analytics for real-time adaptation. These systems would automatically adjust caption presentation based on learners' real-time proficiency (e.g., listening comprehension and vocabulary mastery), guiding them to adopt the most contextually appropriate caption strategies for their current learning stage. This aligns with the perspective articulated by Little and McDaniel (2015), who emphasized that future research should aim not only to identify individual learners' strategy differences for outcome prediction, but more importantly, to systematically design interventions that encourage the adoption of contextually optimal strategies across different proficiency levels.

6. Conclusions

In conclusion, this study provides further evidence that proficiency level significantly influences how learners process captioned videos. Low-proficiency learners predominantly rely on captions as a scaffold for comprehension, while high-proficiency learners demonstrate a greater ability to integrate auditory and textual cues. Additionally, video complexity (the presence of multiple speakers) can increase reliance on captions, particularly for low-proficiency learners.

This study contributes to the cognitive applications in education by operationalizing "desirable difficulties" (Bjork & Bjork, 2011) in multimedia instruction. For low-proficiency learners, captions serve as a scaffold that balances challenge and support. Conversely, high-proficiency learners' preference for auditory input (as evidenced by higher listening scores) suggests that reducing caption reliance may introduce beneficial difficulties, potentially enhancing real-time auditory processing. The findings also demonstrate that contextual factors (e.g., video complexity) dynamically modulate the balance between challenge and support. Future research should investigate the development of listening abilities across learners of varying proficiency levels under different captioning modalities, while also

exploring dynamic systems that adapt captions in real time based on learners' evolving linguistic profiles and contextual demands.

Supplementary Materials: The supporting information can be downloaded at https://www.mdpi.com/article/10.3390/bs15040542/s1.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study in accordance with Article 32 of the "Measures for Ethical Review of Human Life Science and Medical Research" issued by the Department of Science and Technology Education of the National Health Commission of the People's Republic of China.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy.

Acknowledgments: I would like to express my sincere gratitude to Yeldham for his invaluable guidance and support rendered throughout the preparation of this thesis.

Conflicts of Interest: The author declares no conflicts of interest.

References

Bjork, R. A., & Bjork, E. L. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Caimi, A. (2006). Audiovisual translation and language learning: The promotion of intralingual subtitles. *The Journal of Specialized Translation*, *6*(1), 86.

Chai, J., & Erlam, R. (2008). The effect and the influence of the use of video and captions on second language learning. *New Zealand Studies in Applied Linguistics*, *14*(2), 25–44.

Cohen, L., Manion, L., & Morrison, K. (2002). *Research methods in education*. Routledge.

Danan, M. (2004). Captioning and subtitling: Undervalued language learning strategies. *Meta*, *49*(1), 67–77. [CrossRef]

Field, J. (2008). Bricks or mortar: Which parts of the input does a second language listener rely on? *TESOL Quarterly*, *42*(3), 411–432. [CrossRef]

Field, J. (2019). *Second language listening: Current ideas, current issues*. Cambridge University Press.

Jarvella, R. J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, *10*(4), 409–416. [CrossRef]

Kennedy, A., Heller, D., Pynte, J., & Radach, R. (Eds.). (2000). *Reading as a perceptual process*. Elsevier.

Krashen, S. (1982). *Principles and practice in second language acquisition*. Pergamon.

Larson-Hall, J. (2015). *A guide to doing statistics in second language research using SPSS and R*. Routledge.

Little, J. L., & McDaniel, M. A. (2015). Some learners abstract, others memorize examples: Implications for education. *Translational Issues in Psychological Science*, *1*(2), 158. [CrossRef]

Mayer, R. E. (Ed.). (2005). *The Cambridge handbook of multimedia learning*. Cambridge University Press.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology*, *93*(1), 187. [CrossRef]

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81. [CrossRef]

Pannatier, M., & Bétrancourt, M. (2024). Learning from academic video with subtitles: When foreign language proficiency matters. *Learning and Instruction*, *90*, 101863. [CrossRef]

Patterson, A. (2019). Exploring paused transcription to assess L2 listening comprehension utilizing Rasch measurement. *Shiken*, *23*(2), 1–18.

Pujolà, J.-T. (2002). CALLing for help: Researching language learning strategies using help facilities in a web-based multimedia program. *ReCALL*, *14*, 235–262. [CrossRef]

Rayner, K., Rotello, C. M., Stewart, A. J., Keir, J., & Duffy, S. A. (2001). Integrating text and pictorial information: Eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied*, *7*(3), 219. [CrossRef] [PubMed]

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. [CrossRef]

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. [CrossRef]

Sydorenko, T. (2010). Modality of input and vocabulary acquisition. *Language Learning and Technology*, *14*, 50–73.

Taylor, G. (2005). Perceived processing strategies of students watching captioned video. *Foreign Language Annals*, *38*(3), 422–427. [CrossRef]

Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. *Quarterly Journal of Experimental Psychology Section A*, *57*(1), 165–182. [CrossRef]

Vandergrift, L. (2004). 1. Listening to learn or learning to listen? *Annual Review of Applied Linguistics*, *24*, 3–25. [CrossRef]

Vanderplank, R. (1988). The value of teletext sub-titles in language learning. *ELT Journal*, *42*(4), 272–281. [CrossRef]

Vanderplank, R. (2016). *Captioned media in foreign language learning and teaching: Subtitles for the deaf and hard-of-hearing as tools for language learning*. Springer.

Vanderplank, R. (2019). 'Gist watching can only take you so far': Attitudes, strategies and changes in behaviour in watching films with captions. *The Language Learning Journal*, *47*(4), 407–423. [CrossRef]

Winke, P., Gass, S., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern Language Journal*, *97*(1), 254–275. [CrossRef]

Yeldham, M. (2016). Second language listening instruction: Comparing a strategies-based approach with an interactive, strategies/bottom-up skills approach. *Tesol Quarterly*, *50*(2), 394–420. [CrossRef]

Yeldham, M. (2017). Techniques for researching L2 listeners. *System*, *66*, 13–26. [CrossRef]

Yeldham, M. (2018a). The influence of formulaic language on L2 listener decoding in extended discourse. *Innovation in Language Learning and Teaching*, *12*(2), 105–119. [CrossRef]

Yeldham, M. (2018b). Viewing L2 captioned videos: What's in it for the listener? *Computer Assisted Language Learning*, *31*(4), 367–389. [CrossRef]

Yeldham, M. (2024). How do second language learners go about their listening when they view captioned videos? Replication studies of Taylor (2005), Winke et al. (2013) and Rodgers and Webb (2017). *Language Teaching*, *57*(4), 575–587. [CrossRef]

Yue, C. L., Bjork, E. L., & Bjork, R. A. (2013). Reducing verbal redundancy in multimedia learning: An undesired desirable difficulty? *Journal of Educational Psychology*, *105*(2), 266–277. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Metacognition About Collaborative Learning: Students' Beliefs Are Inconsistent with Their Learning Preferences

Yunfeng Wei *, Nicholas C. Soderstrom, Michelle L. Meade and Brandon G. Scott

Department of Psychology, Montana State University, P.O. Box 173440, Bozeman, MT 59717-3440, USA; nicholas.soderstrom@montana.edu (N.C.S.); mlmeade@montana.edu (M.L.M.); brandon.scott2@montana.edu (B.G.S.)

* Correspondence: yunfeng.wei@montana.edu; Tel.: +1-406-994-6052

Abstract: Collaboration plays an important role in educational contexts. However, little is known about students' metacognitive beliefs about collaboration. The present study used an online survey to investigate students' beliefs toward group study/recall, their studying preferences, strategies they use when studying individually and in groups, and important characteristics of their group members. Results indicate that, although students generally perceive collaboration as beneficial, they prefer individual study, indicating that their beliefs are inconsistent with their learning preferences. Students report social learning as the primary reason for collaborative benefits but prefer to study alone to minimize distraction and increase personal accountability. Further, they use different strategies when studying individually or in a group. When studying individually, students most frequently report re-reading their notes. However, when studying in groups, students most frequently use strategies emphasizing collaboration and interaction, such as quizzing each other. Also, students prefer to work with group members who are focused, motivated, and hard working. Students' beliefs, preferences, and favored characteristics of group members are related to their frequency of using study groups. Students' metacognitive beliefs about collaboration have implications for theories of self-regulated learning and better use of collaboration in educational contexts.

Keywords: metacognition; collaboration; strategy; collaborative memory; learning

1. Introduction

Students often work together in educational settings. For example, they collaborate on group projects and group discussions, and they rely on study groups to prepare for exams. Such group activities are common in education, yet the effects of collaboration on cognition are mixed [1,2]. Importantly, there is limited research on students' metacognitive beliefs and preferences regarding collaboration. Metacognitive beliefs are an important aspect of learning because students rely on their metacognition to monitor their learning and select appropriate strategies [3]. In the present study, we examine students' perceptions of collaborative study, collaborative testing, their use of study groups, the strategies they use when studying individually and in groups, and which characteristics of their fellow group members students believe are most important for successful collaboration.

2. Collaborative Cognition

Many researchers encourage instructors to use collaboration in educational settings because learning in groups can improve students' motivation, engagement in classes, reasoning, and problem-solving skills [4]. Other researchers have claimed that students who recall information in groups are subsequently better able to recall information individually, a phenomenon called post-collaborative benefits [5–7]. Consequently, many cognitive and educational researchers regard collaborative learning as a powerful pedagogical tool and, therefore, recommend its use in educational contexts.

In spite of its many benefits, collaboration is not always beneficial in terms of memory performance. First, individuals recalling information with other group members may suffer from collaborative inhibition, or worse recall in groups [7,8]. Second, individuals recalling with others may learn misinformation from group members, which is called the social contagion effect [9,10]. Finally, recalling with other individuals can lead to forgetting. Research on socially shared retrieval-induced forgetting demonstrates that listening to someone recall information can lead to forgetting related information [11].

Given the mixed results of collaboration on learning and memory, it is important to understand student perceptions about collaboration. Previous metacognitive research in collaboration has focused on perceptions of group-level processes such as coordinating information and expertise, delegating tasks, developing strategies, etc. [12]. In contrast, the current study is focused on understanding student perceptions of how collaboration influences their learning and memory during study groups and how such perceptions relate to their study behaviors and preferences. Further, given that there are situations in which collaboration is relatively more beneficial or harmful, it is important to understand students' beliefs about the characteristics of successful collaborative groups.

3. Metacognition

Metacognition refers to people's understanding of their own cognitive processes, as well as one's knowledge and understanding of how learning and memory operate [13,14]. Metacognitive research has demonstrated that students rely on metacognition to regulate their learning, and metacognition can directly affect subsequent behavior [15]. For example, individuals' judgments of learning (JOL)—that is, their subjective assessments of their own learning—are linked to self-regulated learning: As judgments of learning become more accurate, students make better study choices [16]. Students' metacognitive beliefs about collaboration, then, have important implications for theories of self-regulated learning because they can inform if and how students rely on others to regulate their learning.

Given the importance of metacognition to self-regulated learning, it is noteworthy that students' metacognitive judgments do not always correspond to their cognitive performance [17,18]. For example, Blasiman et al. [19] demonstrated that students' intended study behaviors may not correspond to their actual study behaviors by showing that students intended to study twice as long as they did. The dissociation between metacognition and cognition is especially relevant to the current study because students' beliefs about study groups may or may not relate to their actual use of study groups and/or how study groups influence their learning.

Theories of self-regulated learning also suggest that students rely on metacognitive beliefs to select study strategies [3]. Again, there is a dissociation between metacognition and cognition in the strategies students use to study, as students often prefer to use non-optimal strategies. For example, Karpicke et al. [20] surveyed college students to investigate the strategies students prefer when studying. They found that students most frequently re-read notes to prepare for exams [21,22]. Further, students are rarely aware of the usefulness of testing as a powerful learning strategy, though the testing effect—the memorial benefit of retrieving information from memory—has been replicated many times in a wide variety of contexts [23,24]. Together, this research suggests that students rely on their metacognitive beliefs about learning to select strategies they believe will benefit their learning in individual contexts. Importantly, little is known about the strategies used in collaborative groups, and it remains unknown if and how students' preferred strategies differ across individual and collaborative study contexts.

4. Students' Perceptions of Collaboration

To our knowledge, only a handful of studies have examined participants' metacognitive beliefs about collaboration. In the current study, we replicate and extend previous research to examine the following factors that we believe can best inform theory and practical applications: (1) the relationship between students' beliefs about collaboration,

their actual use of collaborative groups, and their study preferences; (2) the distinction between collaborative study and collaborative recall; (3) a direct comparison of individual strategies and strategies used in collaborative groups; (4) characteristics associated with effective group members; and (5) differences in metacognitive beliefs between students who frequently and rarely use study groups.

4.1. Beliefs, Frequency of Collaboration, and Preferences

Several studies have surveyed students' beliefs about collaboration and their use of collaborative groups, e.g., [25,26]. These studies demonstrate that students generally believe that recalling in groups is more effective than recalling individually (see too related research showing that individuals view collaboration positively, regardless of whether or not performance is better in the collaborative group [27] or worse in the collaborative group [28].

However, the literature on students' actual use of collaborative groups is more mixed. Wissman and Rawson [25] reported that 41% of students spend at least one hour per week in collaborative groups, and McCabe and Lummis [29] reported that 78% of students study in a group at least once per semester. In contrast, Rybczynski and Schussler [26] reported that students rarely used study groups (55% never used study groups, and 22% tried but quit over the course of the semester). Students' preferences for individual or group study are also inconsistent. McCabe and Lummis reported that nearly half of students prefer to study alone (47.6%) rather than study collaboratively, while Rybczynski and Schussler reported that only a small proportion of students (5 out of 246) prefer to study alone, but students became more likely to prefer to study alone and less willing to participate in collaborative study as the semester progressed. Considered together, existing research suggests that students generally believe that collaboration is beneficial, but it is unclear how frequently they engage in collaborative learning or if they prefer collaborative learning over individual learning.

In the current study, we directly test the relationship between beliefs, frequency, and preference within a single study. Consistent with previous findings demonstrating that metacognition is often inconsistent with cognitive performance [18,19], we hypothesized that students would have a positive attitude toward study groups but rarely use study groups and prefer to study alone, indicating their beliefs are inconsistent with their actual and preferred behaviors.

4.2. Group Recall and Group Studying

Wissman and Rawson [25] asked students questions specific to collaborative study and collaborative testing. Interestingly, only 25% thought studying in a group was more beneficial than studying alone, but regarding testing specifically, 51% said they learn more when testing in a group vs. alone. When asked why testing is more effective in groups, the most common explanation (22% of participants) was that collaborative testing helps students figure out what they do not know because they cannot look at the answer. These results highlight the important distinction between collaborative studying and collaborative testing and, importantly, demonstrate that students differentiate study and testing dimensions of collaboration. However, this study did not fully code participants' qualitative responses, and so it remains unknown why students believe collaboration has different effects on studying and testing.

In the current study, we ask students open-ended questions about why they believe studying/recalling in groups/individually is more effective. Our hypotheses here are exploratory, as we are seeking to further characterize students' metacognitive beliefs. Nonetheless, we hypothesized that when students choose to study or recall in groups and study or recall individually, their underlying reasons would be different.

4.3. Strategies Used When Studying Individually and in Groups

Previous research has established that undergraduate students report re-reading notes as a preferred strategy to study for exams individually [20–22]. Relatively less is known about students' preferred strategies when studying for exams in groups. Wissman and Rawson [25] found that 79% of students reported quizzing each other. Rybczynski and Schussler [26] reported the most common strategy used in collaborative groups (98% of students) was discussing or clarifying notes with group members. Finally, McCabe and Lummis [29] reported that the most frequently used strategy in collaborative groups was to ask questions.

In these studies, participants were presented with several strategies and asked to check all the strategies they use during collaborative learning [25,26] or to rate how frequently they engaged in each strategy [29]. The three studies presented different strategy options for students to select from, so it is difficult to directly compare results. However, using an open-ended question asking which strategies students used in groups, McCabe and Lummis reported that the most frequently listed group strategy was to work on practice problems, followed by discussing course materials, quizzing or testing each other, and asking questions, results that are broadly consistent with the forced choice data. Importantly, however, these studies did not ask students about their preferred strategies in groups, and none of these studies included an individual comparison condition.

In the current study, we extend this research to ask open-ended questions about the strategies students use during collaboration, and critically, we also ask them to rank order their preferred/most used study strategies [20]. Further, we include a comparison of individual study strategies to determine how the frequency/rating of strategies differs across individual and group study. Because students highlight collaboration in study groups [25,26,29], we hypothesized that students use similar strategies when studying in groups compared to studying individually but that they prioritize more interactive strategies when studying in groups because, unlike studying individually, studying in groups affords such interactive strategies.

4.4. Characteristics of Other Group Members

Another exploratory element of this study was to examine what characteristics students believe are important of their group members. Rybczynski and Schussler [26] provided a qualitative analysis of student beliefs about collaboration. Students reported that group composition is important, equal contribution from group members is vital, study groups often lack focus and are not productive, and that social learning has inherent value. The surveys conducted throughout the semester were administered immediately following class exams, and so students' perceptions of any collaboration prior to the exam were likely influenced by their exam performance. Nonetheless, these data suggest that students' metacognitive beliefs about collaborative study groups depend at least partially on how the group performs. This is relevant to our interest in better understanding what characteristics students believe contribute to successful study groups.

In the current study, we replicate and extend the results of Rybczynski and Schussler [26]. Consistent with their results showing that students believe that focus, equity, and productivity are the major concerns of students studying in groups, we hypothesized that such characteristics would also be important to our participants. However, given our open-ended format, we were able to identify other characteristics that students found important of their group members.

4.5. Differences Between Students Who Frequently and Rarely Use Study Groups

As previous studies show [25,26,29], some students frequently use study groups, while some others rarely or never use study groups. Further, there is some evidence that students who frequently use study groups have different metacognitive beliefs compared to those who rarely use study groups. Wissman and Rawson reported that 71% of students who rarely studied in groups (less than 1 h per week) believed individual study was more

beneficial, while only 40% of students who frequently studied in groups (more than 1 h per week) believed individual study was more beneficial. Likewise, McCabe and Lummis reported that students who frequently use group study use more effective strategies during group study. In the current study, we replicate and extend these results to examine how experience influences metacognitive beliefs about collaboration. That is, besides surveying students' (1) beliefs, (2) preferences, (3) strategies, and (4) group member features, we split participants into those who frequently use groups and those who rarely use groups to further examine how students' previous study group experiences are related to their metacognition about study groups.

Finally, we note that this study was conducted during the COVID-19 pandemic. Research suggests that social distancing and isolation during the pandemic significantly impacted students' well-being [30] and students' learning environments as many classes moved online [31]. Further, Rosenfield et al. [32] argue that group processes and interpersonal relationships are among the psychological phenomena most likely to show lasting impacts from the COVID-19 pandemic. As such, this study is unique in examining student perceptions of collaboration during the pandemic and serves as a time point that can help future researchers examine if/how student perceptions about collaborative learning change in the coming years as any effects of the pandemic settle.

5. Method

5.1. Participants

Two hundred and two students participated in the online survey as a part of the Introduction to Psychology course credits. The mean age of participants was 20.10 years, and the majority of participants were first-year students (121 first year, 46 second year, 24 third year, nine fourth year, and two others). The sample included 81 males, 118 females, and three third gender, and most of the participants were White (93%). The sample size in the present study was determined in accordance with previous studies that examined the strategies students use [20,21,25,26]. Blasiman et al. [19] recommended 200 or more participants for survey research of student strategies; our sample size exceeded this recommendation.

The data were collected in the spring semester of 2021 during the COVID-19 pandemic. At this time, Montana State University was in the middle of its second semester of blended learning (students alternated between attending class in person and online), and students had experienced disruptions to their personal and academic lives. As a result, the sample used in the current study offers a unique and important perspective on students' metacognitive beliefs about collaboration during the pandemic.

5.2. Materials and Procedure

Survey questions were developed locally or adapted from previous research [20,26,33] (see Appendix A for the full survey). The survey included questions regarding participants' beliefs about studying in groups versus studying individually and their study behaviors, including the following: "How often do you use study groups?"; do you prefer to study for exams in a group or individually?; which do you think is more effective—studying individually or studying in groups?; and which do you think is more effective—recalling individually or recalling in groups? Asking participants which is more "effective" is consistent with past research on the strategies that students perceive as most effective for learning [33]. For these questions, we used a forced-choice format rather than a Likert scale to avoid any neutral or ambiguous responses. That is, we wanted participants to directly compare these two methods and select which one they thought was more effective. We also asked participants to answer why they thought individual or collaborative study/recall was more effective, and why they preferred to study individually or collaboratively. The "Why" questions were open-ended and helped verify that participants interpreted the question as effective for learning. Furthermore, participants were asked to list strategies they used when studying individually and in groups, from the most frequently used to the least frequently used. Participants also listed characteristics of other group members they

thought were important for the group to be successful and were asked to explain why such characteristics are important. Finally, the survey included an open-ended question asking for any additional thoughts about study groups.

6. Results

6.1. Beliefs, Frequency of Collaboration, and Preferences

We used chi-square tests to investigate students' beliefs and preferences about studying. Firstly, when asked if they believed whether studying in a group or studying individually was more effective, an equal number of participants chose each option. As shown in Figure 1, there was no significant difference in the number of students who believed that studying in groups or individually was more effective, $\chi^2(1, n = 202) = 0.02$, $p = 0.89$. However, when asked about recalling in groups, more students believed that it was more effective to recall in a group than to recall information individually, $\chi^2(1, n = 202) = 6.42$, $p = 0.01$.

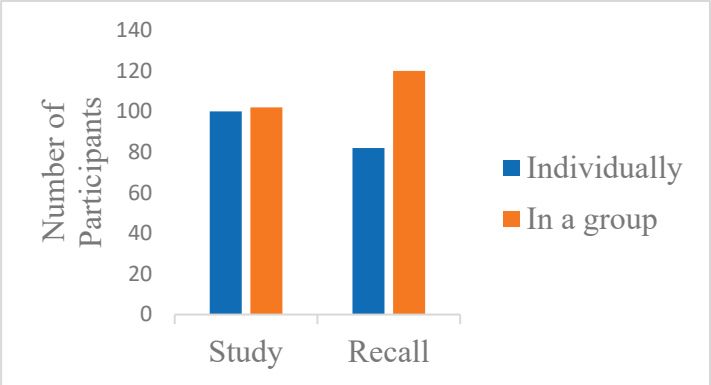


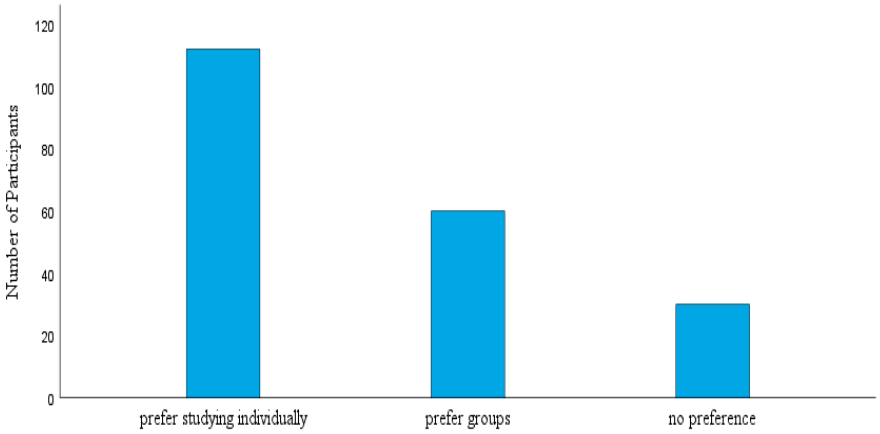
Figure 1. Which do you think is more effective—studying in groups or studying individually? Which do you think is more effective—recalling in groups or recalling individually?

Surprisingly, we found that although most students believed group study to be equally effective and group recall to be more effective than individual recall, most students preferred to study individually for their exams (see Figure 2a), $\chi^2(2, n = 202) = 51.13$, $p < 0.001$.

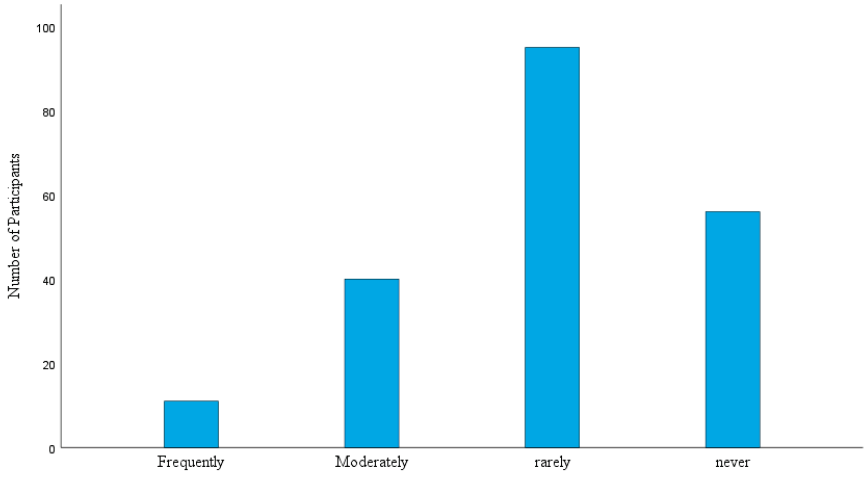
Lastly, when students were asked whether they used study groups, many of them reported that they rarely use study groups (see Figure 2b), $\chi^2(3, n = 202) = 72.89$, $p < 0.001$. These findings suggest that students believe recalling in study groups is beneficial, but they prefer to study individually and rarely use study groups, indicating their actual and preferred study behaviors are not consistent with their beliefs.

6.2. Group Recall and Group Studying

We then analyzed the responses to the questions "Why do you think studying in groups/individually is more effective?" and "Why do you think recalling in groups/individually is more effective?". Participants' responses were coded into different categories adapted from Rybczynski and Schussler [26] or constructed locally by following the best practices for qualitative data coding outlined by Syed and Nelson [34]. For example, many students wrote "because I want to learn from other students"; we categorized these responses as social learning. The first 30 percent of responses were coded by two persons, and interrater reliability was calculated using Cohen's kappa in SPSS to ensure that the reliability of each column was above 60% [35]. After interrater reliability was high, one of the coders finished the rest of the responses.



(**a**)



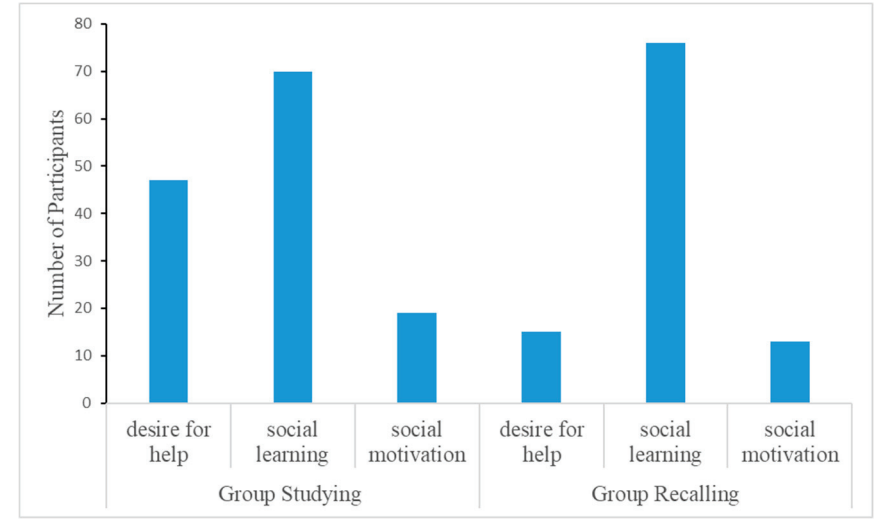
(**b**)

Figure 2. (**a**) Do you prefer to study for exams in a group or individually? (**b**) How often do you use study groups?

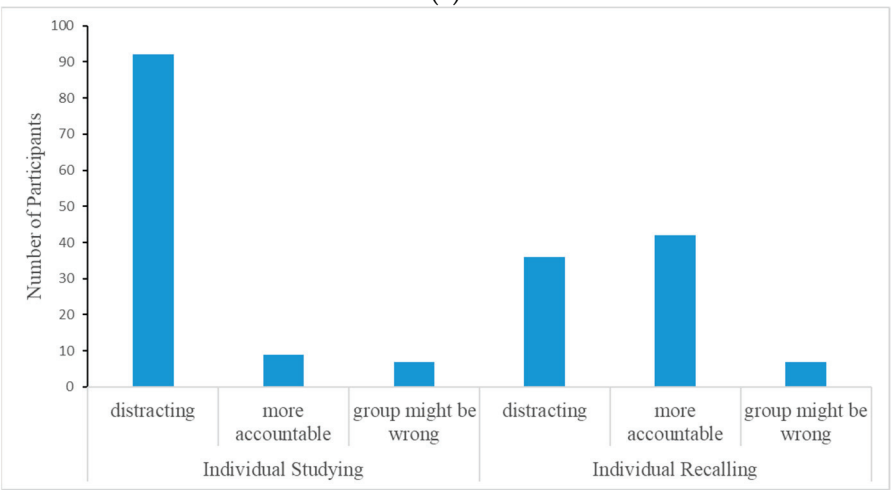
As shown in Figure 3a,b, there were three main reasons why participants preferred to study/recall in groups: (1) desire for help—to obtain help from other members, (2) social learning—to learn from other members, and (3) social motivation—work hard due to social pressure. Figure 3a,b also shows the three main reasons why participants prefer to study/recall individually: (1) less distraction—groups are distracting, (2) more accountable—to make sure individuals master the materials instead of hearing answers from other members, and (3) misinformation—individuals were worried about learning incorrect information from other members. These responses underscore that participants interpreted our questions of "effective" strategies as strategies that are effective for learning.

We compared the distribution of their responses in Figures 3a and 3b using chi-square tests for independence. We report *Cramér's V* as a measure of effect size. When interpreting *Cramér's V* for the chi-square test for independence when $df^* = 1$, a small effect size is 0.1, a medium effect size is 0.3, and a large effect size is 0.5 or greater [36,37]. We found that the distribution of studying and recalling in groups was different, $\chi^2(2, 240) = 13.87$, $p < 0.001$, *Cramér's V* = 0.24, and the distribution of studying and recalling individually was different, $\chi^2(2, 193) = 43.73$, $p < 0.001$, *Cramér's V* = 0.48. Specifically, the main reason for choosing to both study/recall in groups is social learning. However, compared to recalling in groups, more participants choose to study in groups because they desire help. Regarding the primary reasons for studying/recalling individually, most students want

to study individually because study groups are distracting. Many students also want to recall individually because groups are distracting, but this is not the dominant reason for recalling individually. Rather, many students want to recall individually because they are more accountable; they want to ensure they can retrieve information individually.



(**a**)



(**b**)

Figure 3. (**a**) Distribution of responses to why preferring studying in groups/individually. (**b**) Distribution of responses to why preferring recalling in groups/individually.

6.3. Strategies Used When Studying Individually and in Groups

Table 1a,b shows the strategies students listed when studying individually and in groups. Looking first at the strategies reported when students study individually, we found that 69.8% of participants reported they re-read notes and textbooks, and the mean rank was high, indicating a large number of students use re-reading materials frequently. Also, many students (over 20%) reported they use flashcards, do practice problems, do practice recall/Quizlet, use review sheets, re-write notes, and memorize/repeat materials.

Table 1. (**a**) Strategies students use when studying individually. (**b**) Strategies students used when studying in groups.

(**a**)					
Strategy	Percent who list strategy		Percent who rank as #1 strategy		Mean rank
Re-reading notes and textbooks	69.8	(141)	32.7	(66)	1.8
Flashcards	35.6	(72)	10.4	(21)	2.3
Do practice questions	34.1	(69)	11.9	(21)	2.1
Practice recall/Quizlet	31.2	(63)	12.9	(26)	2.0
Watching videos/outside resources	28.7	(58)	5.9	(12)	2.5
Review sheet	27.7	(56)	8.9	(18)	2.3
Re-write notes	20.8	(42)	9.9	(20)	1.8
Memorize/repetition	20.3	(41)	8.9	(18)	2.0
Focus on wrong answers/unclear information	9.9	(20)	0.5	(1)	2.9
Mnemonics/association	6.9	(14)	1.5	(3)	2.4
Highlight (in notes/books)	5.9	(12)	1.5	(3)	2.2
Create a good environment	5.9	(12)	3.5	(7)	1.6
Spaced repetition/interleaving	4.0	(8)	1.0	(2)	2
Think of real life examples	2.5	(5)	0.5	(1)	2.8
(**b**)					
Strategy	Percent who list strategy		Percent who rank as #1 strategy		Mean rank
Teach each other/discussion	44.1	(89)	21.2	(43)	1.7
Quiz each other	40.1	(82)	13.9	(28)	2.0
Ask for help/answers	22.3	(45)	10.4	(21)	1.8
Flashcards	20.8	(42)	9.4	(19)	1.9
Compare answers/notes	20.3	(41)	7.4	(15)	2.0
Re-reading notes/textbooks	20.3	(41)	5.9	(12)	2.3
Do practice problems	18.3	(37)	9.4	(19)	1.8
Make review sheets/study guide	18.3	(37)	9.9	(20)	1.6
Practice recall/Quizlet	8.9	(18)	2.5	(5)	2.3
Focus on wrong answers	7.9	(16)	3.0	(6)	1.9
Re-write notes	7.9	(16)	1.0	(2)	2.8
Watch videos/ask professors/outside resources	6.4	(13)	1.5	(3)	2.8
Group focuses on one topic	4.0	(10)	1.0	(4)	2
Mnemonics/association	3.5	(7)	1.0	(2)	2.3

Note. The numbers in parentheses reflect the raw participant numbers.

Turning next to the strategies reported when students study in groups, over 40% of students reported that they teach each other/discuss and quiz each other in groups. Over 20% of students reported that they ask for help/answers and compare notes with group members. Notably, these strategies were never listed as strategies used by individuals. However, the rest of the strategies listed for group study were also listed for individual study, but they were listed less frequently. Therefore, these results demonstrate that the strategies students use when studying in groups/individually are generally consistent but that participants prioritize interactive strategies.

6.4. Characteristics of Other Group Members

Table 2 displays the important characteristics of other group members listed by participants from the most reported to the least reported (more than five cases). More than 50% of participants claimed that "focused/task-oriented" is an important characteristic of other group members. This corresponds to the previous findings that lacking focus is one of the biggest concerns of groups [26]. In addition, more than 30% of participants reported the characteristics responsible/hard working (39.1%) and motivated/desire to participate (35.1%). More than 20% of participants thought being cooperative/flexible was

important (23.8%). More than 10% of participants hoped their group members could have some positive attributes, such as being helpful/patient (19.8%), kind/considerate (13.9%), and easy-going/fun/talkative (10.9%). Building on these findings, many students think characteristics that can help complete tasks are important, and some students also think having positive traits is important for their group members.

Table 2. Important characteristics of other group members.

Characteristics	Percent Who List Strategy	
Focused/task oriented	56.4	(114)
Responsible/hard-working/reliable	39.1	(79)
Motivated/desire to participate	35.1	(71)
Flexible/receptive/cooperative	23.8	(48)
Helpful/patient	19.8	(40)
Kind/considerate	13.9	(28)
Easy-going/talkative/fun	10.9	(22)
Communication skills/organized	8.9	(18)
Have similar goals	5.4	(11)
Humble/willing to make mistakes	5.4	(11)
Similar knowledge or skills	5.0	(10)
Similar/equal contribution	4.5	(9)
Leadership	4.0	(8)
Skeptical/critical	3.0	(6)

Note. The numbers in parentheses reflect the raw participant number.

6.5. Differences Between Students Who Frequently and Rarely Use Study Groups

We split our participants into two groups—"frequently" or "moderately" used groups and "rarely" or "never" used groups. By comparing these groups' studying and recalling choices, preferences, preferred strategies when studying individually and in groups, and favored characteristics of other group members, we can investigate how previous experience with groups is related to students' metacognitive beliefs. Again, we compared the distribution of their responses using chi-square tests for independence. Because this required six chi-square tests, we adopted 0.008 as the significant p value instead of 0.05 in order to correct for type 1 error.

Table 3 shows beliefs and preferences for students who frequently/moderately use study groups and students who rarely/never use study groups. More students who frequently/moderately use study groups think studying in groups is more beneficial, but more students who rarely/never use study groups think studying individually is more beneficial, $\chi^2(1, 202) = 15.74$, $p < 0.001$, *Cramér's V* = 0.28. This result indicates that students' previous group study experience is related to their studying beliefs and qualifies our previous finding that there was no significant difference in studying beliefs when we combined students who often use and rarely use study groups. Importantly, however, previous group experience had no impact on the perceived benefits of collaborative recall; more students believe recalling in groups to be more effective regardless of whether they previously use study groups or not, $\chi^2(1, 202) = 3.54$, $p = 0.06$, *Cramér's V* = 0.13. We also found that previous study group experiences are related to study preferences. Specifically, as shown in Table 3, more students in the frequently/moderately used group prefer to study for exams in groups, while more students in the rarely/never group prefer to study for exams individually, $\chi^2(1, 172) = 31.39$, $p < 0.001$, *Cramér's V* = 0.43.

In addition, we examined whether group use experiences are related to strategies students use when studying in groups and individually. We only included the top six strategies for analysis because we needed to ensure that there were enough cases in each cell for analysis. Previous group use experiences do not correlate with strategies students use in groups, $\chi^2(5, 340) = 3.44$, $p = 0.63$, *Cramér's V* = 0.10 and individually, $\chi^2(5, 460) = 2.24$, $p = 0.82$, *Cramér's V* = 0.07, suggesting most frequently used strategies are not influenced by students' group experiences.

Table 3. Beliefs and preferences for students who frequently/moderately use study groups vs. students who rarely/never use study groups.

		Frequently/Moderately Group	Rarely/Never group
Beliefs about studying	Study in groups is more effective	38	64
	Study individually is more effective	13	87
Beliefs about recalling	Recalling in groups is more effective	36	84
	Recalling individually is more effective	15	67
Studying preference	Study in groups for exams	32	28
	Study individually for exams	15	97
Favored characteristics of other group members	Focused	26	114
	Motivated	17	71
	Reliable	22	79
	Flexible	23	48
	Helpful	22	40

Note: some students reported no studying preference.

Last, we examined whether group use experiences are related to students' favored characteristics of other group members. Again, we picked the 5 top characteristics because we needed to make sure there were enough free report responses in each cell for the chi-square analysis. More students in the rarely/never group reported characteristics that can help achieve their goals, such as focused, motivated, and reliable $\chi^2(4, 352) = 22.68$, $p < 0.001$, *Cramér's V* = 0.25. This indicated that more students who rarely or never use study groups like group members who can achieve goals, but students who frequently or moderately use study groups also hope other members are helpful and flexible.

7. Discussion

The present study investigated students' beliefs about study groups and further characterized the reasons for their beliefs. We were also interested in how students' beliefs corresponded to why they want to study/recall in groups, strategies they use individually and in groups, and important characteristics of other group members.

7.1. Beliefs, Frequency of Collaboration, and Preferences

In the current study, more students believed *recalling* in groups is more effective than recalling individually; however, there was no significant difference in the number of students who believed *studying* in groups was more effective than studying individually. These findings indicate that more students have a positive attitude toward group recalling, which is consistent with previous findings [25–27,29]. However, when studying for exams, more students prefer to study individually [29]. Additionally, nearly half of our participants rarely use study groups, and many of them never use study groups. Thus, metacognition about group studying is not consistent with real or preferred studying behaviors [19]. In general, students think study groups are beneficial, but they prefer to study individually.

Previous experience in collaborative groups influenced these conclusions to some extent. Those frequently using study groups believe group study to be more effective and prefer to study for exams in groups. In contrast, previous experience had no impact on students' beliefs that collaborative recall was effective. More generally, students' beliefs and preferences are related to their frequencies of using study groups.

7.2. Group Recall and Group Studying

Wissman and Rawson [25] reported that individual studying is thought to be more effective, but group testing is regarded as more helpful. We partially replicated their findings by showing that there was no significant difference in the number of students who

believe studying individually or in groups is more effective, but more students believe recalling in groups is more effective. In the current study, we found that nearly 50% of students believe studying in groups is more effective, whereas only 21% of Wissman and Rawson's participants thought so. We used an online chi-square calculator [38] to compare the frequency data in the current study with those in Wissman and Rawson and found that the percentage in the current study is significantly higher ($p < 0.00$). Given that Wissman and Rawson did not ask participants why they thought studying individually was more effective, the difference could be ascribed to students' increasing awareness of the advantages of group study. Additional explanations for the discrepancy between studies include the use of different samples and possible time of measurement effects; COVID-19 could impact student perceptions of learning and collaboration.

We also extended Wissman and Rawson's [25] findings by asking students why they believe studying/recalling individually/in groups is more effective using open-ended questions. Specifically, compared to studying in groups, fewer students desire others' help when recalling in groups. The predominant reason for studying individually is that groups are distracting, while the main reason for recalling individually is to make sure individuals can recall the information in the future instead of group members telling them the answer. These data are relevant to theories of self-regulated learning because they characterize students' beliefs about why they seek out collaborative study/recall or individual study/recall to regulate their learning. Further, because students believe studying and recalling are different processes, experimenters and instructors should give students more specific instructions in the future instead of using a general term, such as "learning".

7.3. Strategies Used When Studying Individually and in Groups

Replicating previous research [20], most students like to re-read their notes and textbooks when studying individually. In addition, many students also reported that they do practice problems, use flashcards, make review sheets, and practice recall, all of which is consistent with Karpicke et al.'s study. The only difference in the current study was that many students reported that they would like to watch videos and use some other online tools, such as YouTube. This could be attributed to the development and accessibility of online technologies, as well as the reliance on online courses due to COVID-19. The most novel findings from our strategy questions involve collaborative groups. Importantly, when studying in groups, more students prefer to teach each other/discuss, quiz each other, and ask for help. That is, when studying in groups, students are more likely to use strategies emphasizing collaboration and interaction. These findings are conceptually consistent with previous studies [25,26,29] in demonstrating that collaborative groups frequently quiz each other and compare and share notes.

Interestingly, we found that previous experience in collaborative groups had no impact on preferred strategies. This finding does not align with McCabe and Lummis [29] claiming that students with more study group experiences use more effective strategies. The disagreement between findings can be ascribed to the design of our study. In their study, the authors listed all the strategies, and students could rate the frequency of each one, but we asked our participants to list strategies freely. It is possible that students judge the frequency of using various strategies differently when they are given choices or are asked to list freely.

7.4. Characteristics of Other Group Members

More than half of students hope their group members are focused, which supports the findings of Rybczynski and Schussler [26] that individuals think lack of focus is one of the biggest problems of study groups. This finding is also conceptually consistent with collaborative memory research that demonstrates group composition can influence group performance [39,40]. Many participants reported that they did not want their groups to go off-topic. Additionally, many students prefer task-oriented characteristics, including being motivated and responsible. Some students also prefer their group members to be

talkative, cooperative, and helpful. These characteristics can provide individuals with a supportive and encouraging environment in which they collaborate. Therefore, many participants think their group members should have some characteristics that are important to achieve their goals, and some hope their group members are supportive and positive. It is worth noting that students frequently/moderately using study groups also emphasize the helpfulness and flexibility of their group members, while those rarely/never using study groups only highlight goal-relevant characteristics.

7.5. Theoretical and Practical Implications

This study has important practical and theoretical implications for self-regulated learning. Self-regulated learning requires students to plan for their learning, monitor their on-going learning, and adjust their study behavior if necessary. There are many theories and models regarding the relationship between these elements of self-regulated learning [16,41]. According to the results of the present study, students may consider the unique strategies afforded by collaboration when regulating their own learning. For example, when studying in groups, the current participants prioritized interactive strategies that allow them to clarify information and quiz each other to test how much information each group member has mastered. Our participants also reported nuanced beliefs about how collaboration can help or hurt their learning during study and retrieval and how group composition affects learning. Given these findings, there is a great need for theories of metacognition to include collaboration as a potential factor that may influence how students' monitor and regulate their own learning. Toward this end, the present study can serve as a starting point for investigating how students' beliefs about collaborative study, collaborative recall, and group composition affect their use of study groups, their learning preferences, and how they monitor and adapt their learning processes.

It is important that educators understand students' concerns about study groups so they can help students better use study groups in educational contexts. Also, students can learn that they may be underusing study groups. Specifically, although students believe study groups are beneficial—and in many cases, study groups *are* beneficial—they typically adhere to the traditional studying method. Students may, therefore, consider using study groups more frequently and examine their effectiveness. Taken altogether, our findings can be used to equip educators and students with knowledge about how to better incorporate study groups into their own educational practices.

8. Limitations and Future Directions

We acknowledge several limitations in the current study. The sample was collected from one university at a single point in time. As such, it is important to replicate the results across a wider range of participants to demonstrate generalizability, especially when COVID-19 influenced the way students are educated to a large extent [30–32]. Future research can examine students' metacognition about collaboration again after the pandemic to determine any lasting effects of the pandemic on metacognitive beliefs about collaboration. Further, the questionnaire was aimed at collaborative learning generally rather than focusing on the role of collaborative learning for specific class topics. Further research is necessary to determine if the results presented here generalize across different content and materials. Also, we asked students which strategies are "effective" [33], which could be interpreted in different ways (e.g., improving their grades, assisting in their long-term learning, or enhancing their short-term performance). Answers to our "why" questions suggested that students interpret it as effective for learning. However, future research can examine more precisely which aspects of learning students believe are most influenced by collaboration.

Further research is necessary to follow up on and extend some of the key findings from the present study. The results of this study suggest that there is a discrepancy between students' beliefs about study groups and students' learning preferences. That is, students seem to realize that study group is beneficial but prefer to study individually for exams.

Building on this inconsistency between metacognition and learning behavior, further research can probe why this discrepancy occurs and whether an intervention of some kind can help students voluntarily adopt the use of study groups to improve their performance. Students' beliefs about studying in groups or individually are related to their group study experiences. Future research can investigate why previous study group experiences are related to studying beliefs and preferences. Specifically, researchers can examine if previous experiences are the cause or the results of their beliefs and preferences. Or there may be a third factor that can influence students' group use experiences and beliefs. For example, extraverted students may believe study groups are more beneficial and may even benefit more from studying groups than their introverted counterparts.

9. Conclusions

In conclusion, this study investigated students' metacognition about group study and showed that students' learning activity is inconsistent with their attitudes towards group studying. Namely, students infrequently use study groups, choosing instead to study individually, even though they recognize the general benefits of group study. Their beliefs and preferences are related to their frequencies of using study groups. We think such a discrepancy is important, from both a theoretical and practical standpoint, and is a worthy topic for future research, such as factors that discourage students from study groups, the application of collaborative recall in classrooms, and instructors' impact on adjusting students' metacognition and preferences.

Author Contributions: Conceptualization, Y.W., N.C.S. and M.L.M.; methodology, Y.W., N.C.S. and M.L.M.; validation, Y.W., M.L.M. and B.G.S.; formal analysis, Y.W., M.L.M. and B.G.S.; investigation, M.L.M.; resources, Y.W., N.C.S. and M.L.M.; data curation, Y.W. and M.L.M.; writing—original draft preparation, Y.W.; writing—review and editing, N.C.S., M.L.M. and B.G.S.; visualization, Y.W.; supervision, M.L.M.; project administration, M.L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Montana State University (protocol code MM121720-EX and date of approval 17 December 2020).

Informed Consent Statement: All participants read and signed the consent form at the beginning of the survey.

Data Availability Statement: Data for this experiment are publicly available on OSF: https://osf.io/3fx4d/?view_only=6510a49c0d874f7999d70c19c56a869c (accessed on 1 July 2022).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

1. How often do you use study groups?
   Frequently (once or more per week)
   Moderately (once or twice per month)
   Rarely (once or twice per semester)
   Never
2. Which do you think is more effective—studying in groups or studying individually?
   Why do you think it is more effective?
3. Which do you think is more effective-- recalling (remembering) information in groups or recalling information individually?
   Why do you think it is more effective?
4. Do you prefer to study for exams in a group or individually?
   Why do you prefer to study this way?
5. List the strategies you use when studying individually and rank the strategies in terms of how frequently you use them (the first strategy you list should be the one you use most

frequently; the last strategy you list should be the one you use least frequently). (e.g., 1, XXX 2, XXX 3, XXX 4, XXX 5, XXX)
6. List the strategies you use when studying in a group and rank the strategies in terms of how frequently you use them (the first strategy you list should be the one you use most frequently; the last strategy you list should be the one you use least frequently). (e.g., 1, XXX 2, XXX 3, XXX 4, XXX 5, XXX)
7. What characteristics of other group members do you think are important in order for study groups to be productive? (e.g., XXX, XXX, XXX, XXX. . .)
   Why do you think these characteristics are important?
8. Do you have any other thoughts regarding study groups that you would like to share?
9. Your age (in years): _____
10. Gender:
   Male
   Female
   Prefer not to say
   Other: _____
11. Current year in school:
   First year college
   Second year college
   Third year college
   Fourth year college
   Other: _____
12. Ethnicity:
   Hispanic or Latino
   Not Hispanic or Latino
13. Race:
   American Indian/Alaska Native
   Asian
   Black/African American
   Native Hawaiian/Other Pacific Islander
   White
   Prefer not to say
   Other:_____

References

1. Rajaram, S.; Pereira-Pasarin, L. Collaborative Memory: Cognitive Research and Theory. *Perspect. Psychol. Sci.* **2010**, *5*, 649–663. [CrossRef] [PubMed]
2. Pociask, S.; Rajaram, S. The effects of collaborative practice on statistical problem solving: Benefits and boundaries. *J. Appl. Res. Mem. Cogn.* **2014**, *3*, 252–260. [CrossRef]
3. Bjork, R.A.; Dunlosky, J.; Kornell, N. Self-regulated learning: Beliefs, techniques, and illusions. *Annu. Rev. Psychol.* **2013**, *64*, 417–444. [CrossRef] [PubMed]
4. Persellin, D.; Daniels, M.; Reder, M. *A Concise Guide to Improving Student Learning*; Stylus Publishing, LLC.: Sterling, VA, USA, 2015.
5. Wei, Y.; Charbonneau, B.Z.; Meade, M.L.; Hutchison, K.A. Examining the time course of post collaborative benefits across word lists and prose passages. *Mem. Cogn.* **2024**, 1–12. [CrossRef] [PubMed]
6. Blumen, H.M.; Rajaram, S. Influence of re-exposure and retrieval disruption during group collaboration on later individual recall. *Memory* **2008**, *16*, 231–244. [CrossRef]
7. Rajaram, S. Collaborative Inhibition in Group Recall: Cognitive Principles and Implications. In *Collaborative Remembering: Theories, Research, and Applications*; Oxford University Press: Oxford, UK, 2018.
8. Weldon, M.S.; Bellinger, K.D. Collective memory: Collaborative and individual processes in remembering. *J. Exp. Psychol. Learn. Mem. Cogn.* **1997**, *23*, 1160–1175. [CrossRef]
9. Meade, M.L.; Roediger, H.L. Explorations in the Social Contagion of Memory. *Mem. Cogn.* **2002**, *30*, 995–1009. [CrossRef]
10. Gabbert, F.; Wheeler, R. Memory conformity following collaborative remembering. In *Collaborative Remembering: Theories, Research, and Applications*; Meade, M.L., Harris, C.B., Van Bergen, P., Sutton, J., Barnier, A., Eds.; Oxford University Press: Oxford, UK, 2018.

11. Hirst, W.; Yamashiro, J. Social aspects of forgetting. In *Collaborative Remembering: Theories, Research, and Applications*; Meade, M.L., Harris, C.B., Van Bergen, P., Sutton, J., Barnier, A., Eds.; Oxford University Press: Oxford, UK, 2018.
12. Biasutti, M.; Frate, S. Group metacognition in online collaborative learning: Validity and reliability of the group metacognition scale (GMS). *Educ. Tech Res. Dev* **2018**, *66*, 1321–1338. [CrossRef]
13. Flavell, J.H. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *Am. Psychol.* **1979**, *34*, 906–911. [CrossRef]
14. Koriat, A. Metacognition and Consciousness. In *The Cambridge Handbook of Consciousness*; Zelazo, P., Ed.; Cambridge University Press: New York, NY, USA, 2007; pp. 289–325.
15. Nelson, T.O.; Narens, L. Metamemory: A theoretical framework and new findings. In *The Psychology of Learning and Motivation*; Bower, G.H., Ed.; Academic Press: New York, NY, USA, 1990; pp. 1–45.
16. Dunlosky, J.; Metcalfe, J. *Metacognition*; Sage Publications, Inc.: Thousand Oaks, CA, USA, 2009.
17. Rhodes, M.G. Judgements of Learning: Methods, data, and Theory. In *The Oxford Handbook of Metamemory*; Dunlosky, J., Tauber, S.K., Eds.; Oxford University Press: Oxford, UK, 2016.
18. Soderstrom, N.C.; Yue, C.L.; Bjork, E.L. Metamemory and education. In *The Oxford Handbook of Metamemory*; Dunlosky, J., Tauber, S.K., Eds.; Oxford University Press: Oxford, UK, 2015.
19. Blasiman, R.; Dunlosky, J.; Rawson, K. The what, how much, and when of study strategies: Comparing intended versus actual study behaviour. *Memory* **2017**, *25*, 784–792. [CrossRef]
20. Karpicke, J.D.; Butler, A.C.; Roediger, H.L., III. Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory* **2009**, *17*, 471–479. [CrossRef]
21. Gurung RA, R. How Do Students Really Study (and Does It Matter)? *Teach. Psychol.* **2005**, *32*, 239–241.
22. Hartwig, M.K.; Dunlosky, J. Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychon. Bull. Rev.* **2012**, *19*, 126–134. Available online: https://www.proquest.com/scholarly-journals/study-strategies-college-students-are-self/docview/1021977684/se-2 (accessed on 5 October 2024). [CrossRef]
23. Roediger, H.L., III; Karpicke, J.D. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychol. Sci.* **2006**, *17*, 249–255. [CrossRef]
24. Rawson, K.A.; Dunlosky, J. When is practice testing most effective for improving the durability and efficiency of student learning? *Educ. Psychol. Rev.* **2012**, *24*, 419–435. [CrossRef]
25. Wissman, K.T.; Rawson, K.A. How do students implement collaborative testing in real-world contexts? *Memory* **2015**, *24*, 223–239. [CrossRef]
26. Rybczynski, S.M.; Schussler, E.E. Student use of out-of-class study groups in an introductory undergraduate biology course. *CBE Life Sci. Educ.* **2011**, *10*, 74–82. [CrossRef]
27. Cortright, R.N.; Collins, H.L.; Rodenbaugh, D.W.; DiCarlo, S.E. Student retention of course content is improved by collaborative-group testing. *Adv. Physiol. Educ.* **2003**, *27*, 102–108. [CrossRef]
28. Whillock, S.R.; Meade, M.L.; Hutchison, K.A.; Tsosie, M.D. Collaborative inhibition in same-age and mixed-age dyads. *Psychol. Aging* **2020**, *35*, 963–973. [CrossRef]
29. McCabe, J.A.; Lummis, S.N. Why and how do undergraduates study in groups? *Scholarsh. Teach. Learn. Psychol.* **2018**, *4*, 27–42. [CrossRef]
30. Dimmock, J.; Krause, A.E.; Rebar, A.; Jackson, B. Relationships between social interactions, basic psychological needs, and wellbeing during the COVID-19 pandemic. *Psychol. Health* **2021**, *37*, 457–469. [CrossRef]
31. Besser, A.; Flett, G.L.; Zeigler-Hill, V. Adaptability to a sudden transition to online learning during the COVID-19 pandemic: Understanding the challenges for students. *Scholarsh. Teach. Learn. Psychol.* **2022**, *8*, 85–105. [CrossRef]
32. Rosenfeld, D.L.; Balcetis, E.; Bastian, B.; Berkman, E.T.; Bosson, J.K.; Brannon, T.N.; Burrow, A.L.; Cameron, C.D.; Chen, S.; Cook, J.E.; et al. Psychological Science in the Wake of COVID-19: Social, Methodological, and Metascientific Considerations. *Perspect. Psychol. Sci.* **2022**, *17*, 311–333. [CrossRef]
33. McCabe, J.A. Metacognitive awareness of learning strategies in undergraduates. *Mem. Cogn.* **2011**, *39*, 462–476. [CrossRef]
34. Syed, M.; Nelson, S.C. Guidelines for Establishing Reliability When Coding Narrative Data. *Emerg. Adulthood* **2015**, *3*, 375–387. [CrossRef]
35. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
36. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Hillsdale, N.J., Ed.; L. Erlbaum Associates: Mahwah, NJ, USA, 1988.
37. Gravetter, F.J.; Wallnau, L.B. *Essentials of Statistics for the Behavioral Sciences*, 8th ed.; Wadsworth/Cengage: Belmont, CA, USA, 2014.
38. Chi-Square Test Calculator. (2024, September 10). Available online: https://www.socscistatistics.com/tests/chisquare2/default2.aspx (accessed on 5 October 2024).
39. Hart, K.M.; Meade, M.L. Social contagion of memory and the role of self-initiated relative judgments. *Acta Psychol.* **2021**, *212*, 103189. [CrossRef]

40. Numbers, K.T.; Barnier, A.J.; Harris, C.B.; Meade, M.L. Ageing stereotypes influence the transmission of false memories in the social contagion paradigm. *Memory* **2019**, *27*, 368–378. [CrossRef]
41. Winne, P.H. Modeling self-regulated learning as learners doing learning science: How trace data and learning analytics help develop skills for self-regulated learning. *Metacognition Learn.* **2022**, *17*, 773–791. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Investigating the Effectiveness of Self-Regulated, Pomodoro, and Flowtime Break-Taking Techniques Among Students

Eva J. C. Smits [1,*], Niklas Wenzel [2] and Anique de Bruin [2]

[1] Faculty of Psychology and Neuroscience, Maastricht University, 6229 ER Maastricht, The Netherlands
[2] Department of Educational Research & Development, School of Health Professions Education, Faculty of Health, Medicine, and Life Sciences, Maastricht University, 6211 LK Maastricht, The Netherlands; niklas.wenzel@maastrichtuniversity.nl (N.W.); anique.debruin@maastrichtuniversity.nl (A.d.B.)
* Correspondence: ejc.smits@alumni.maastrichtuniversity.nl

Abstract

Effective break taking during study sessions is crucial for maintaining performance, especially in self-regulated learning settings where students plan their own tasks. However, research on break taking in these contexts is limited. This study investigates the effect of self-regulated, Pomodoro, and Flowtime breaks on subjective study experiences, task completion, and flow. Ninety-four university students participated in an online intervention that provided instruction on how to take breaks in a 2 h authentic study session. In the self-regulated break condition (n = 25), students decided when and how long to take breaks. In the Pomodoro condition (n = 36), students took 5 min breaks after every 25 min of studying. In the Flowtime group (n = 33), participants decided when to take breaks; however, the break duration was determined based on prior study duration. Results showed that Pomodoro breaks led to a faster increase in fatigue, and Pomodoro and Flowtime breaks led to a faster decrease in motivation compared with self-regulated breaks; however, these differences did not result in overall differences in fatigue or motivation levels between conditions. Similarly, no differences were found in productivity levels, task completion, and flow. Future research should further examine these techniques by including variables like personality and mental effort.

Keywords: self-regulated break taking; pomodoro; flowtime; study experience; self-regulated learning

1. Introduction

Effective learning is a demanding and effortful process that can lead to feelings of stress and decreases in productivity levels and well-being (Fritz et al., 2013; Ginns et al., 2023). Recovery from the demands of study tasks is essential for maintaining high levels of performance during the day (Phan & Beck, 2023). Taking breaks from study tasks allows for the opportunity to recover personal resources (e.g., motivation and energy levels), which are necessary for an effective study session (Trougakos & Hideg, 2009). Taking breaks may therefore counteract the negative effects of prolonged task engagement and may help to complete tasks more efficiently (Li et al., 2020; Lyubykh et al., 2022; Phan & Beck, 2023). The specific timing and duration of breaks may have differential effects on the recovery of these resources (Lim & Kwok, 2016; Trougakos & Hideg, 2009). Moreover, taking effective breaks can enhance the study experience of students and may improve their well-being and mental health (Lee et al., 2021; Phan & Beck, 2023). Despite the benefits of effective

break taking, some challenges students face in self-regulating these breaks are still not fully understood.

One of these challenges is how students can effectively regulate their breaks, especially in contexts outside the classroom, where students must manage their learning independently. Self-regulated learning (SRL) is defined by Pintrich (2000) as "an active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation, and behavior, guided and constrained by their goals and the contextual features in the environment" (p. 453). SRL is thus an independent and self-directed process in which students are fully responsible for regulating their learning (Goetz et al., 2013). In these settings, students are therefore required to make decisions about when, where, and how to study. A lot of students struggle with self-regulating their learning because of difficulties in attention retention, motivation, and lack of metacognitive knowledge (Biwer et al., 2023a; Schloemer & Brenan, 2006; Zimmerman, 2002; Zimmerman & Moylan, 2009).

Deciding to engage in study tasks involves the decision to forgo other potentially rewarding activities, such as socializing or going outside. According to the opportunity costs model of Kurzban et al. (2013), this trade-off constitutes opportunity costs. The model suggests that the perceived costs of a task are influenced by the appeal of other available opportunities and the perceived benefits of investing effort in the task at hand (Kurzban et al., 2013). For example, studying for an exam may be perceived as less effortful if a student expects it will lead to a higher grade. Taking a break from study tasks provides an opportunity to engage in alternative activities, potentially reducing the perceived effort and restoring motivation (Dora et al., 2022). In this way, break taking may support learning by restoring personal resources. Break taking can, but is not necessarily, part of self-regulated learning, as students' break taking may impact the achievement of their learning goals (Lyubykh et al., 2022). There are different techniques for regulating breaks during self-regulated learning. Some students may plan breaks in advance, using techniques that externally regulate breaks, while others rely on subjective cues such as fatigue, productivity, and perceived effort and progress productivity to decide when to disengage from learning tasks (Bjork & Bjork, 2020; de Bruin et al., 2023; Grahek et al., 2022).

Given the challenges students face while self-regulating their learning, it is important to understand how students can improve their study technique by effectively regulating their break taking to enhance their study experience. Therefore, additional research on the effectiveness of different break-taking techniques in self-regulated learning settings is needed. The following section outlines different break regulation techniques and their potential implications for student learning.

2. Types of Break Regulation Techniques

2.1. Self-Regulated Breaks

In self-regulated break taking, students take breaks whenever they want. The decision to take a break is based on a subjective feeling and is therefore not planned. The term "self-regulated breaks", as used in this paper, refers to taking breaks whenever a student feels like it. Students also determine for themselves when and how long to take a break. A commonly described disadvantage of self-regulated breaks is that they can increase the cognitive load by inducing an additional self-monitoring load. The mental effort that is used for this cannot be used for studying, which can make studying more cognitively demanding (Lee, 2022).

In addition, deciding yourself when to take a break is a form of self-interruption. Self-interruptions are interruptions that are initiated internally (Katidioti et al., 2016). Contrarily, external interruptions are interruptions that occur due to something outside of the self

(Abad et al., 2018; Katidioti et al., 2016). In self-regulated breaks, individuals interrupt their study tasks in response to an internally driven desire for a break. Research has shown that self-interrupted tasks are completed more slowly than externally interrupted tasks (Katidioti et al., 2016). Due to these self-interruptions, self-regulated break taking may reduce motivation and productivity, which may prolong the time needed to complete study tasks (Katidioti et al., 2016).

To minimize the additional mental load and self-interruptions, various break-taking techniques have been developed that are thought to enhance effectiveness; however, research on these techniques in self-regulated learning settings is limited.

2.2. The Pomodoro Technique

The Pomodoro technique is a well-known technique, as it is frequently promoted through online platforms (e.g., Scroggs, n.d.; Staff, 2025). When using the Pomodoro technique, students study in blocks of 25 min with a 5-min break in between. These blocks are repeated four times before taking a longer break of fifteen to thirty minutes (Sheldon & Wigmore, 2022). In this case, breaks are externally regulated by a timer (Dizon et al., 2021). Even though the break duration and timing are regulated externally, students still have to self-regulate their own learning and regulate their breaks by setting the timer. Systematic breaks like Pomodoro are thought to postpone distractions and facilitate motivation and concentration. In addition, Pomodoro breaks are thought to improve productivity and reduce mental fatigue. The Pomodoro technique might specifically diminish procrastination because it requires students to split big tasks into smaller sections, which can lead to a higher degree of task completion (Almalki et al., 2020). The Pomodoro technique is therefore often recommended online, even though it has received surprisingly little attention in education research.

Biwer et al. (2023b) conducted a study about systematic break taking and its effect on mental effort, task completion, and study experiences. In this study, three independent conditions were included, one self-regulated break-taking group and two systematic break-taking groups. One condition was similar to Pomodoro, and the other systematic break-taking group used even smaller study blocks of 12 mininute with a 3 mininute break in between. They found that students in both systematic break conditions reported being more concentrated and motivated, and perceived learning tasks to be less difficult, than in the self-regulated group (Biwer et al., 2023b).

Even though the Pomodoro technique is thought to have a lot of advantages, one critique is that this technique can disrupt the so-called flow state. This is a state in which someone is highly focused and productive. In this state, people usually do not feel time passing and feel absorbed in their tasks (Csikszentmihalyi, 1975). For a student to experience flow, complete involvement in the current task as well as the loss of awareness of everything outside of the task is necessary (Willekes, 2014). The flow state is very fragile; for example, this state can already be disrupted when a small distraction is noticed (Willekes, 2014). A flow state can therefore already be interrupted by a thought about a grade or a small mistake (Zielke et al., 2023). Because Pomodoro uses a timer during study time and is very strict about when to take a break, this technique can disrupt the flow state (Thompson et al., 2021).

2.3. The Flowtime Technique

A break-taking technique that might overcome this problem is the Flowtime technique. The technique is also known as the "Flowmodoro" technique, since it is a variation of the Pomodoro technique (Shen, 2023). In the Flowtime technique, students study for as long as they can focus and take a break whenever they feel the need. The duration of the break

depends on how long they have worked on their study tasks. For example, if someone studies for 25 min or less, they will receive a 5 min break, like in the Pomodoro technique. However, if someone decides to study for a longer period of time, they will receive a longer break (Lau, 2023). This technique combines systematic elements of the Pomodoro method and self-regulatory elements of the self-regulated breaks. This combination may, on the one hand, lead to a disadvantage compared with the Pomodoro technique since students still need to self-regulate when to take a break, which may increase cognitive load, which can make studying more demanding (Lee, 2022). However, since students do not have to determine the length of their breaks themselves, cognitive demands are still not as high as with completely self-regulated breaks. On the other hand, the self-regulated element in the Flowtime technique leads to a considerable advantage over the Pomodoro technique. Since the Flowtime technique does not use a timer during study time, the flow state will not be interrupted, and it therefore resolves the critique of the Pomodoro technique (Oliveira dos Santos et al., 2018).

3. The Present Study

This study aims to examine the differences between the self-regulated, Pomodoro, and Flowtime break techniques and their effectiveness during self-regulated learning in higher educational settings. The relation between break-taking technique and motivation, fatigue, productivity, task completion, and flow will be investigated during an authentic self-regulated learning session. Therefore, the present study is centered around the following question: "What is the effect of the type of break-taking technique (self-regulated, Pomodoro, or Flowtime breaks) on motivation, productivity, fatigue, task completion, and flow state?" To test this research question, several hypotheses were formulated:

Hypothesis 1.

(a) *The Pomodoro and Flowtime conditions will show higher levels of motivation compared with the self-regulated break condition.*
(b) *The Pomodoro and Flowtime conditions will show higher levels of productivity compared with the self-regulated break condition.*
(c) *The Pomodoro and Flowtime conditions will show lower levels of fatigue compared with the self-regulated break condition.*

Hypothesis 2. *The Pomodoro and Flowtime conditions will show higher levels of task completion compared with the self-regulated break condition.*

Hypothesis 3. *The flow state will be greater in the self-regulated break and Flowtime conditions than in the Pomodoro condition.*

Furthermore, this paper examines whether different break-taking techniques influenced changes in subjective study experiences over time and whether individuals exhibited distinct trajectories over time. This is addressed by a second exploratory research question: "How do different break-taking techniques (self-regulated, Pomodoro, and Flowtime breaks) influence the trajectory of students' subjective study experiences (namely motivation, productivity, and fatigue) during a study session, and are there differences in these changes between individuals?"

4. Materials and Methods

4.1. Participants and Design

The sample of this study included 111 psychology students from a Dutch university, which was deemed necessary based on the less advanced analysis method initially planned, as the study was conducted for a bachelor's thesis. Since a more advanced, better-suited analysis method is used in the current paper, the required sample size for achieving sufficient power might have been higher. Ethical approval was obtained from the ethical board of the university (The Ethical Review Committee of Psychology and Neuroscience, ERCPN) with the corresponding reference code ERCPN-274_120_11_2023. Participants were recruited online through the SONA system, which is a system that allows universities to create participant pools and upload studies.

Participants were randomly assigned to one of three conditions: self-regulated, Pomodoro, or Flowtime breaks. All groups received exactly two hours to study and take breaks. The time students studied between one break and the next break is called a "study block". Students in the self-regulated group took self-regulated breaks and were allowed to study and take breaks as they felt like doing. Participants in this group were instructed to study as long as they wanted and afterward take a break as long as they needed. The second group used the Pomodoro technique for taking breaks and was instructed to study in four blocks of 25 min with 5 min breaks in between. To ensure consistency across all groups, participants in this condition did not receive the extended break after the fourth study block. Providing these students with a longer final break would have exceeded the two-hour limit. Moreover, since participants in the other conditions did not consistently end their sessions with a break, this could have biased the results of this study. The last group used the Flowtime technique and was therefore instructed to study as long as they could focus, and they were allowed to take a break if they could not focus any longer, as in the self-regulated group. The duration of the break these participants received was dependent on the duration of the previous study block. Students received a 5-min break after studying for 25 min or less, and an 8-min break if they studied between 25 and 50 min. If students studied for more than 50 min after the last break, they received a 10-min break. The precise instructions that participants received per condition can be examined in Appendix A.

4.2. Measures

Data collection and survey distribution occurred via the Qualtrics survey software. Participants could access the Qualtrics survey via SONA. The full questionnaire with all measures described below is provided in Appendix B.

4.2.1. Subjective Study Experience

Motivation, productivity, and fatigue were measured in a similar way as in the study by Biwer et al. (2023b). This was accomplished with questions like, "How motivated do you feel right now?". In the study by Biwer et al. (2023b), these questions were answered on a 5-point Likert scale. However, due to a human error, the questions were answered by the participants on a 4-point Likert scale in this study. Every participant answered one item per measure (i.e., motivation, productivity, and fatigue) before starting the study session, after every break, and after the two hours of study.

4.2.2. Task Completion

Task completion was also measured in a similar way as in the study by Biwer et al. (2023b). Participants indicated at the beginning of the study session what tasks they were planning to work on. Participants had to indicate at least one task and could indicate up

to eight tasks. After the 2 h study session, the participants indicated for each task to what extent they had completed the task on a scale from 0 to 100 percent. To compute the total task completion, the mean of these completion percentages across all tasks was calculated.

4.2.3. Passed Time and Break Duration

Time was measured using the timing option in Qualtrics. This option measures the time a student is at a particular part of the questionnaire, i.e., at the "Time for study" or at the "Time for a break" part. The participants were instructed to keep the Qualtrics survey open during the whole study period. The participants could thus use their laptops for their study tasks but could leave the questionnaire open in the background. By doing this, the Qualtrics software automatically measured the time students spent studying and taking breaks.

4.2.4. Flow State

Lastly, a flow questionnaire was used to measure to what extent participants experienced they were in a flow state during the 2 study hours. For this, the Flow short scale (FSS) was used, which is a 10-item questionnaire designed to assess the state of flow. It utilizes a 7-point Likert scale ranging from 1 (not at all) to 7 (very much). This scale was designed in German and translated by Rheinberg et al. (2023) in English. This version of the questionnaire is reliable and is supported by evidence of validity for use in flow state research with a Cronbach's alpha of 0.90 (Rheinberg et al., 2023). The English version was used for this study, and participants filled it in after the study session. The total flow state was calculated by adding up the scores on all items.

4.3. Procedure

Data collection for the current study took place between January and March 2024. Figure 1 provides an overview of the procedure of the study. Before signing up, participants were informed that the research focused on break taking among students and that it would take approximately 120 min to complete. Participants were also informed that the study would be carried out online. After signing up, they were given the opportunity to read the information letter with information about the research and procedure and gave informed consent before the start of the study.

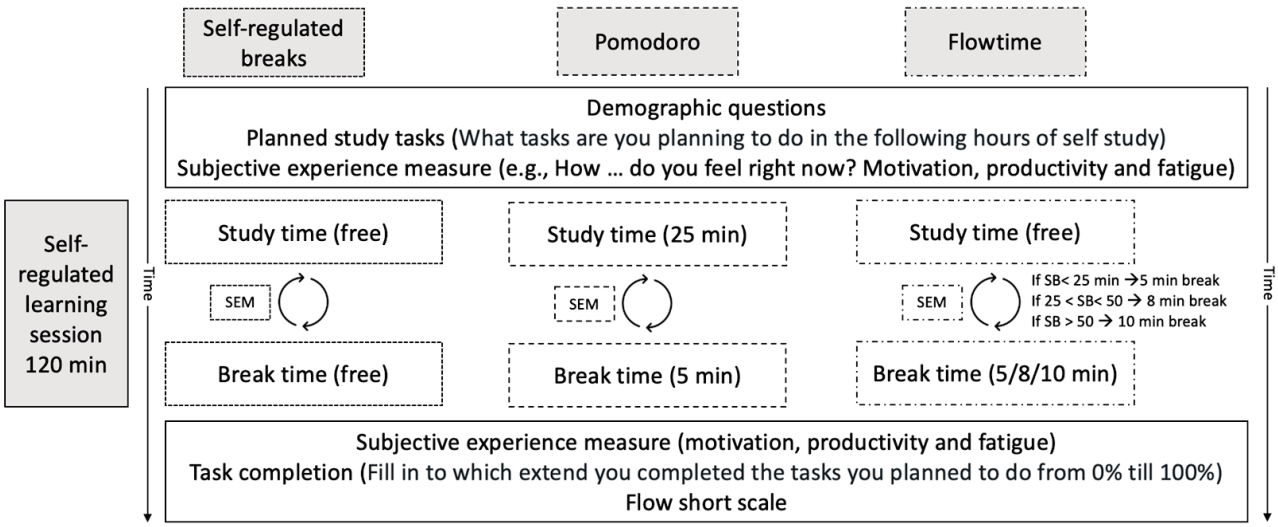


Figure 1. Study procedure for all conditions (self-regulated breaks, Pomodoro, and Flowtime). SEM is the subjective experience measure, SB is study block.

The study relied on an experimental design commonly used in cognitive psychology, but it was conducted in an authentic study setting. Students could engage with their own academic materials with minimal interference from the research. Participants first completed a short questionnaire about their current motivation, productivity, and fatigue levels, planned tasks, and some demographic questions. After this, all participants were randomly assigned to one of three conditions. In all three groups, participants were instructed to work on the study tasks that they had intended to work on in the following study block of two hours. Subsequently, the three groups received distinct instructions about taking breaks; the precise instructions that the participants received are provided in Appendix A.

The self-regulated group was instructed to press a button in the Qualtrics environment when they started and ended their break. For the Pomodoro group, the questionnaire automatically switched between study time and break time. An audio signal informed participants when to take a break and when to continue studying. The Flowtime group was asked to press a button when starting their break, and a sound was played to inform them that they should start studying again.

In addition to hearing a sound, every participant was instructed on the screen what they should be doing at a specific time. The screen all participants saw while studying stated, "Start to study!" and the break screen stated, "Time for a break!" In the Pomodoro and Flowtime groups, the break screen also indicated the total duration of the current break.

After every break, participants in all conditions were asked to indicate how motivated, fatigued, and productive they felt. Furthermore, immediately after the two hours of self-regulated learning were completed, the participants were directed to a last set of questions about their subjective study experiences, task completion, and flow state. For some participants the questionnaire failed to transfer to the last part of the questionnaire automatically; these participants were asked to click a link that would directly lead to the last set of questions after the two hours had passed.

4.4. Analysis

We tested three main sets of hypotheses regarding the effects of different break-taking techniques on motivation, productivity, fatigue, task completion, and flow. Outcome measures were collected at different time points. Motivation, productivity, and fatigue were assessed repeatedly on a 4-point Likert scale. For these outcome variables, we used cumulative link mixed effects models (CLMMs) to account for the ordinal nature of the data as well as the unbalanced repeated measures design. In contrast, task completion (calculated as a percentage of completed tasks) and flow (assessed using 10 items on a 7-point Likert scale) were measured only once at the end of the study period. For these latter outcomes, we aggregated the repeated measures (using means for the motivational and productivity variables) and applied linear regression analysis.

4.4.1. Preparing the Analysis

Data cleaning and preparation were performed prior to the main analyses to ensure data quality, adherence to study parameters, and suitability for the planned statistical tests. Participants were excluded if their data did not meet the predetermined criteria necessary for valid analysis. Specifically, exclusions were made based on non-adherence to study time parameters. This involved removing individuals whose total recorded time exceeded a cutoff of 130 min (n = 9), thereby excluding prolonged participation potentially indicative of off-task behavior. Additionally, this cutoff allowed us to account for natural inefficiencies that characterize authentic learning environments, such as delayed switches from break to studying and vice versa. Participants were also excluded if their engagement with their

self-selected study tasks was minimal (study time < 10 min, n = 2) or if their break time was excessive (total break time > 110 min, n = 1), as this indicated insufficient engagement to provide meaningful data relevant to our research questions. Furthermore, exclusions were necessary when participants' data were unsuitable for the core analyses central to our study. This included participants with incomplete repeated measures data (n = 2), which prevented their inclusion in longitudinal analyses, and those who took no breaks during the study session (n = 2). As our analyses focused specifically on the characteristics and effects of breaks, data from this latter group could not contribute to testing these specific hypotheses. Finally, one participant was identified as a statistical outlier whose overall data pattern suggested they were not representative of the target sample or could disproportionately influence results. This participant was initially flagged as an outlier via IQR on age (age = 38) and was excluded because they also showed divergent patterns on multiple key study engagement variables, indicating an anomalous response profile. In total, 17 participants were removed based on these combined criteria related to study adherence, data completeness, analytic requirements, and anomalous patterns, reducing our analysis sample from 111 to 94 participants.

The final sample's characteristics were n = 94, with a mean age of 20.02 years (SD = 1.87) and 87 (93%) women. When broken down by group, the self-regulated group included 25 participants (mean age 19.84, SD = 1.60; women = 22), the Flowtime group consisted of 33 participants (mean age 20.20, SD = 1.98; women = 31), and the Pomodoro group had 36 participants (mean age 20.00, SD = 1.97; women = 34). All variables were re-coded and designated as the appropriate type within R (continuous, factor, etc.) before model estimation.

4.4.2. Analysis Methods

To test our hypotheses, we employed two main modeling strategies that corresponded to the distinct measurement structures in our study. As mentioned above, for the repeatedly measured outcomes (motivation, productivity, and fatigue, each assessed on a 4-point Likert scale), we used CLMMs. This choice was driven by the ordinal nature of the data and the need to account for unbalanced repeated measures within participants, which CLMMs can account for. Starting with fully saturated models that incorporated theoretically relevant predictors and random effects, we iteratively simplified these models based on likelihood ratio tests, information criteria (AIC), and theoretical considerations until arriving at the final, robust specifications. After the final model specification, the hypothesis was tested by testing nested models (with and without the predictor in question) using likelihood ratio tests and information criteria (AIC). All relevant assumptions were checked (see below).

For outcomes that were measured only once as post-measures (task completion and flow), we aggregated repeated measures (using means) where needed and applied linear regression. In these analyses, standard assumptions (normality of residuals, linearity, homoscedasticity, independence, and multicollinearity) were evaluated using both visual and statistical measures (e.g., Q-Q plots, Shapiro–Wilk, Breusch–Pagan, and Durbin–Watson tests with a $p > 0.05$ threshold) to ensure model adequacy. Finally, we tested for potential influential cases by visually inspecting the residual Q-Q plots for each model and by using the interquartile range criteria for continuous variables. We then re-fitted the final models for each hypothesis with these potential influential cases removed. In most cases, this did not lead to a significant difference. The exception to this was the test for H1c, where the removal of two cases at the extreme negative end of the Q-Q plot led to the number of breaks predictor becoming significant.

This general modeling approach—CLMMs for ordinal repeated measures and linear regression for aggregated single-point outcomes—provided a structured and theoreti-

cally informed foundation for testing our hypotheses. Specific details regarding model specifications and assumption diagnostics are provided in the respective results sections.

4.4.3. Model Building and Assumption Checks

For the cumulative link mixed model (CLMM) analyses, we followed a structured model-selection process, beginning with the most complex models deemed theoretically plausible for each hypothesis and aiming to identify the best-fitting, most parsimonious model. We started by determining the most appropriate linking function for the ordinal outcome variable for each hypothesis by comparing probable link functions using AIC and BIC. We compared the logit and the loglog function. The former is often the default choice when it comes to fitting ordinal models and models with a symmetric relationship between predictors and the response variable. We chose to compare this logit function against the loglog function because some of our outcome variables showed a slight tendency toward a left skew, and the loglog function can account for this. Once we established the appropriate linking function, we determined the optimal random effects structure. Our initial complex structure included a random intercept for participant and random slopes for the duration of the last break. We chose this because the duration of the break was one of the variables that could vary between groups, and adding it as a random slope allowed us to assess whether individuals responded differently to the duration of the break per se (beyond the effect of group). We systematically evaluated this more complex model against simpler, nested alternatives (e.g., random intercept only) using likelihood ratio tests (LRTs) for nested models supplemented by AIC and BIC for converging evidence. When criteria conflicted, we favored the more parsimonious model. With the linking function and random effects structure established, we simplified the fixed effects using a backward elimination approach. Starting with the full model containing all theoretically relevant predictors and their interactions, we sequentially tested the removal of terms, beginning with higher-order interactions, using LRTs and changes in AIC/BIC. Non-significant terms were removed unless they were main effects involved in significant or marginally significant interactions. Throughout this simplification, the centered time-passed variable and group variable were always retained to account for the repeated-measures design and due to their centrality to our hypotheses. This iterative process continued until only statistically significant predictors, terms involved in significant interactions, or theoretically essential predictors remained, resulting in the final model used for hypothesis testing.

We used several methods to test the relevant assumption for the CLMMs. A visual inspection of Q-Q plots was used to verify that the random effects approximated a normal distribution. The proportional odds assumption was evaluated via nominal test (using clm() and the nominal_test() function of the ordinal package (v2023.12-4.1; Christensen, 2023) with $p > 0.05$ taken as evidence of no violation. This meant that we had to refit the final model without the random effects structure because the clmm() function was not supported by the nominal_test() function at the time of writing. Importantly, H1b showed a potential violation of the proportional hazards assumption. Visual checks using cumulative probability plots (cumulative probabilities for each level of a given predictor plotted against the outcome thresholds) indicated minor departures for the group and fatigue predictors at the lowest threshold of the outcome variable (productivity). Additionally, the lines of cumulative probabilities across the levels of each predictor tended to converge at the highest level of the outcome variable. Based on these evaluations, we assumed that the proportional hazards assumption is mostly intact but that the findings related to productivity as outcome should be treated with care. Finally, models without the random effects structure were fitted and compared (via likelihood ratio tests and AIC) to ensure that the inclusion of random effects improved the model fit.

For the linear regression models, we conducted standard assumption checks. The normality of individual variables and of residuals was confirmed via visual Q-Q plots and the Shapiro–Wilk test (with $p > 0.05$ indicating acceptable normality). Linearity was checked through residual plots. Homoscedasticity was assessed with both a visual inspection of the residuals and the Breusch–Pagan test (again using $p > 0.05$ as supportive evidence). Independence of residuals was evaluated using the Durbin–Watson test. Multicollinearity was examined by calculating variance inflation factors (VIFs) using the car package. None of these tests indicated any violation of assumptions for H2 (task completion) or H3 (flow). In addition, potential influential cases were identified via visual inspection of residuals and re-fitting the models. In neither case did this change the results significantly.

5. Results

The means and standard deviations of variables measured in the current study can be found in Table 1.

Table 1. Mean and standard deviation of total study and break time; number of breaks; average motivation, productivity, and fatigue levels; number of planned tasks; total task completion; and flow state in the study for all break-taking conditions separately.

Variable	Self-Regulated Breaks		Pomodoro		Flowtime	
	M	*SD*	*M*	*SD*	*M*	*SD*
Total study time	1.71	0.26	1.69	0.03	1.72	0.14
Total break time	0.30	0.26	0.34	0.02	0.30	0.14
Number of breaks	3.04	1.06	4.00	0	2.82	1.65
Average motivation	2.86	0.41	2.73	0.63	2.59	0.57
Average productivity	3.03	0.48	2.86	0.61	2.75	0.61
Average fatigue	2.56	0.50	2.59	0.63	2.64	0.61
Number of planned tasks	3.68	1.93	3.11	1.28	3.73	1.74
Task completion	70.32	20.12	69.41	18.70	67.09	17.50
Flow state	4.22	0.83	4.10	0.74	3.98	0.79

Total study time and total break time in hours; task completion in percentages.

5.1. Motivation

Hypothesis 1a posited that the Pomodoro and Flowtime groups would report higher levels of motivation relative to the self-regulated group. To test this, we fitted a CLMM model predicting motivation scores (assessed on a 4-point Likert scale). The final model included fixed effects for group membership; centered passed time since the start of the study; linear and quadratic orthogonal polynomial terms for lagged fatigue, lagged productivity, and lagged motivation; and the interaction between group membership and centered passed time. Participant-level random intercepts were included to account for the non-independence of repeated measures within individuals. Table 2 summarizes the model estimates.

Model evaluation indicated that while the random intercept variance (0.54, $SD = 0.73$) suggests some individual differences in baseline motivation log-odds, a model without random effects (CLM) provided a slightly better fit according to BIC (846.0 vs. 849.4) and was not significantly worse according to a likelihood ratio test ($\chi^2(1) = 2.46$, $p = 0.117$), indicating that the effect of the duration of the last break before the repeated measurement did not differ significantly between participants. However, the CLMM was retained to explicitly account for the nested data structure inherent in the repeated measures design. Random effects appeared normally distributed based on a Q-Q plot. The final model showed a significantly better fit than a null model including only random intercepts and thresholds ($\chi^2(11) = 89.78$, $p < 0.001$, BIC = 849.4 vs. 874.4). Examination of the proportional

odds assumption using a nominal test on a fixed-effects equivalent model and visual inspection of coefficients across binary splits suggested no major violations, although the linear term for lagged motivation showed some deviation at the lowest threshold, warranting caution in its interpretation.

Table 2. Overview of coefficients for motivation levels compared with self-regulated break taking.

Predictor	OR	Log-Odds	Std. Error	*z* Value	*p*-Value
Pomodoro group	1.55	0.44	0.34	1.28	0.200
Flowtime group	0.98	−0.02	0.35	−0.06	0.953
Time (centered)	1.00	−0.00	0.01	−0.33	0.742
Motivation (linear)	55,710.96	10.93	3.76	2.91	0.004 **
Motivation (quadratic)	8.89	2.19	2.62	0.84	0.403
Productivity (linear)	1060.77	6.97	2.95	2.36	0.018 *
Productivity (quadratic)	0.25	−1.37	2.64	−0.52	0.604
Fatigue (linear)	0.01	−4.97	2.73	−1.82	0.069
Fatigue (quadratic)	0.01	−4.43	2.36	−1.87	0.061
Pomodoro × Time	0.98	−0.02	0.01	−2.65	0.008 **
Flowtime × Time	0.99	−0.01	0.01	−2.06	0.039 *

* $p < 0.05$. ** $p < 0.01$.

Threshold coefficients defining the latent motivation levels were estimated as follows: 1 | 2: −2.87 (*SE* = 0.36), 2 | 3: −0.59 (*SE* = 0.27), 3 | 4: 2.35 (*SE* = 0.31).

The participant-level random intercept variance was 0.54 (*SD* = 0.73), indicating notable variability in baseline motivation between participants.

In terms of hypothesis testing, the group comparisons did not reach statistical significance at the reference point of time passed (the midpoint of the study session). Neither the Pomodoro group (*OR* = 1.55, β = 0.44, *SE* = 0.34, *p* = 0.200) nor the Flowtime group (*OR* = 0.98, β = −0.02, *SE* = 0.35, *p* = 0.953) differed significantly in motivation levels from the self-regulated group. However, significant interactions were found between group membership and the centered time variable. To explore this, we calculated the estimated marginal means and compared the slopes of the three groups. This analysis showed that, compared with the self-regulated group, both the Pomodoro group (*OR* = 0.98, β = −0.02, *SE* = 0.01, *p* = 0.008) and the Flowtime group (*OR* = 0.99, β = −0.01, *SE* = 0.01, *p* = 0.039) showed a significantly steeper decline in the odds of higher motivation as the study session progressed, indicating that the odds of being in a higher level of motivation reduced by ca. 2% and 1% more per minute for the Pomodoro and Flowtime group, respectively. Post hoc analyses confirmed these interaction effects. The analysis of slopes (see Figure 2) indicated that the rate of change in motivation log-odds as time passed was significantly more negative for the Pomodoro group compared with the self-regulated group (Δ_β = 0.01, *SE* = 0.01, *p* = 0.024, Bonferroni-adjusted). The slope for the Flowtime group was also more negative than that of the self-regulated group (Δ_β = 0.02, *SE* = 0.01, *p* = 0.118, Bonferroni-adjusted), but it did not reach statistical significance after Bonferroni adjustment.

Regarding the lagged subjective experience variables, the model included orthogonal quadratic polynomial terms. Higher prior fatigue was associated with lower current motivation log-odds, with significant effects for both the linear (*OR* = 0.01, β = −4.97, *SE* = 2.73, *p* = 0.069) and quadratic terms (*OR* = 0.01, β = −4.43, *SE* = 2.36, *p* = 0.061), indicating a nonlinear relationship with approaching significance at the $p < 0.05$ level. Lagged productivity showed a significant positive linear association (*OR* = 1060.77, β = 6.97, *SE* = 2.95, *p* = 0.018), indicating higher prior productivity predicted higher current motivation log-odds; the quadratic term was not significant (*p* = 0.604). Lagged motivation itself showed a strong positive linear association (*OR* = 55710.93, β = 10.93, *SE* = 3.75, *p* = 0.004), indicating moti-

vational persistence from the previous time point; the quadratic term was not significant ($p = 0.403$).

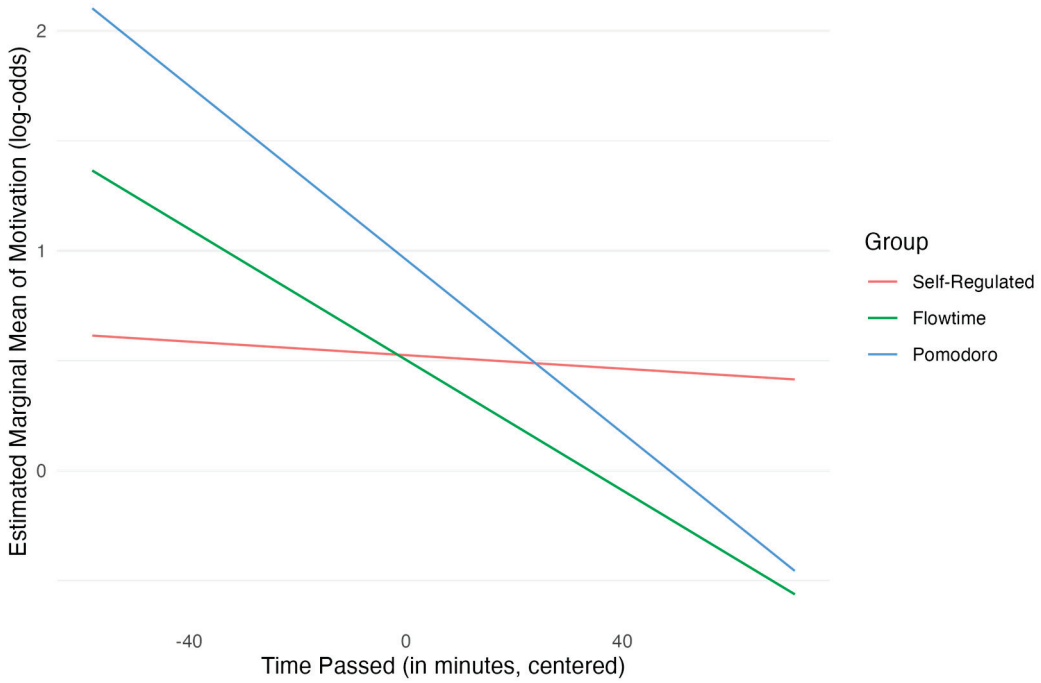


Figure 2. Estimated log-odds of motivation by group over time.

Overall, while the hypothesis of consistently higher mean motivation in the Pomodoro and Flowtime groups compared with the self-regulated group was not supported by the main effects, the significant interaction effects reveal a crucial dynamic: participants using the Pomodoro technique and the Flowtime technique experienced a more pronounced decline in motivation as time elapsed since their last break compared with the self-regulated learners. Furthermore, the significant effects of prior fatigue, productivity, and motivation levels underscore the importance of these dynamic state variables in predicting subsequent motivation during study sessions. It is important to point out that the strong effects of the lagged predictors for productivity and motivation might indicate an issue of (partial) separation, potentially driven by the low number of observations at the extremes of the outcome variable. The effect of these predictors, and especially their magnitude, should therefore be treated with great care.

5.2. Productivity

Hypothesis 1b posited that the Pomodoro and Flowtime groups would report higher levels of productivity relative to the self-regulated group. To test this, we fitted a cumulative link mixed effects model with a logit link function. The final model included fixed effects for group membership; time (centered); orthogonal quadratic polynomial terms for lagged fatigue, lagged productivity, and lagged motivation; and the interaction between group membership and the time variable. Participant-level random intercepts were included to account for repeated measures. Table 3 summarizes the key fixed effects estimates from the model.

The model evaluation revealed that the random intercept variance ($\sigma^2_{intercpt} = 0.16$, $SD = 0.40$) was relatively small. A comparison with a model lacking random effects (CLM) showed the CLM was preferred by AIC (785.7 vs. 787.4) and BIC (840.2 vs. 845.7), and the likelihood ratio test was non-significant ($\chi^2(1) = 0.36$, $p = 0.546$), suggesting the random intercept did not significantly improve model fit, indicating that participants likely do not differ significantly in their level of productivity felt. Despite this, the CLMM was

retained to explicitly model the dependency of repeated measures within individuals. Random effects appeared normally distributed via Q-Q plot inspection. The final model provided a significantly better fit than a null model containing only thresholds and the random intercept ($\chi^2(11) = 72.52$, $p < 0.001$). Tests of the proportional odds assumption (using a fixed-effects equivalent model) indicated no significant violations overall, although lagged fatigue ($p = 0.064$) and lagged motivation ($p = 0.053$) showed marginal significance, suggesting potential deviations, particularly at the lowest threshold where data might be sparser. Interpretation, especially concerning fatigue and motivation effects on transitions into the lowest productivity levels, should proceed with this caution.

Table 3. Overview of coefficients for productivity levels compared with self-regulated break taking.

Predictor	OR	Log-Odds	Std. Error	*z* Value	*p*-Value
Pomodoro group	0.97	−0.04	0.30	−0.12	0.904
Flowtime group	0.75	−0.28	0.30	−0.94	0.348
Time (centered)	1.00	0.00	0.01	0.33	0.739
Motivation (linear)	17,281.66	9.76	2.88	3.39	0.001 ***
Motivation (quadratic)	0.28	−1.29	2.48	−0.52	0.602
Productivity (linear)	116,662.27	11.67	3.43	3.40	0.001 ***
Productivity (quadratic)	19.46	2.97	2.55	1.17	0.244
Fatigue (linear)	0.01	−5.26	2.45	−2.15	0.032 *
Fatigue (quadratic)	0.56	−0.57	2.24	−0.26	0.798
Pomodoro × Time	0.99	−0.01	0.01	−1.44	0.150
Flowtime × Time	0.99	−0.01	0.01	−1.96	0.051

* $p < 0.05$. *** $p < 0.001$.

Threshold coefficients that define the latent productivity categories were estimated as follows: 1 | 2: Estimate = −3.38 (*SE* = 0.35), 2 | 3: Estimate = −1.24 (*SE* = 0.25), and 3 | 4: Estimate = 1.41 (*SE* = 0.25). The participant-level random effects indicated a variance of 0.16 (*SD* = 0.40) for the intercept, indicating relatively low variability in baseline productivity log-odds between participants after accounting for fixed effects. This is in line with the better fit of the CLM model that was discussed above.

In terms of hypothesis testing, the main effects for group comparisons were not statistically significant, indicating that groups did not differ in perceived productivity in the middle of the study session (average of centered time). Neither the Pomodoro group (*OR* = 0.97, β = −0.04, *SE* = 0.29, *p* = 0.904) nor the Flowtime group (*OR* = 0.75, β = −0.28, *SE* = 0.30, *p* = 0.348) differed significantly in productivity log-odds from the self-regulated group at this reference point. The main effect for time was also not significant (*OR* = 1.00, *β* = 0.00, *SE* = 0.01, *p* = 0.739). The interaction between the Pomodoro group and time since the last break was not significant (*OR* = 0.99, β = −0.01, *SE* = 0.01, *p* = 0.150). A marginally significant interaction was observed, however, between the Flowtime group and centered time (*OR* = 0.99, β = −0.01, *SE* = 0.01, *p* = 0.051). This suggests that for participants in the Flowtime group, the log-odds of higher productivity tended to decrease more steeply over time compared with the self-regulated group. Estimated marginal means were calculated and post hoc analyses were performed. Specifically, pairwise comparisons of the group slopes for the log-odds of productivity over time showed that the slope for the Flowtime group was more negative than that for the self-regulated group, consistent with the interaction term. Yet, this difference did not reach statistical significance after Bonferroni adjustment ($\Delta_\beta = 0.01$, *SE* = 0.01, *z* = 1.96, *p* = 0.152). Figure 3 shows the slopes for the three groups across time.

Regarding the lagged state variables, the linear effect of lagged fatigue was significant (*OR* = 0.01, β = −5.26, *SE* = 2.45, *p* = 0.032), indicating that higher prior fatigue was

associated with lower current productivity log-odds. The quadratic term for fatigue was not significant ($p = 0.798$). Both lagged productivity (Linear: $OR = 116{,}662.27$, $\beta = 11.67$, $SE = 3.43$, $p < 0.001$; Quadratic: $p = 0.244$) and lagged motivation (Linear: $OR = 17{,}326.63$, $\beta = 9.76$, $SE = 2.88$, $p < 0.001$; Quadratic: $p = 0.602$) showed strong, significant positive linear associations with current productivity log-odds. This suggests that higher productivity and higher motivation in the previous assessment strongly predicted increased productivity in the current assessment.

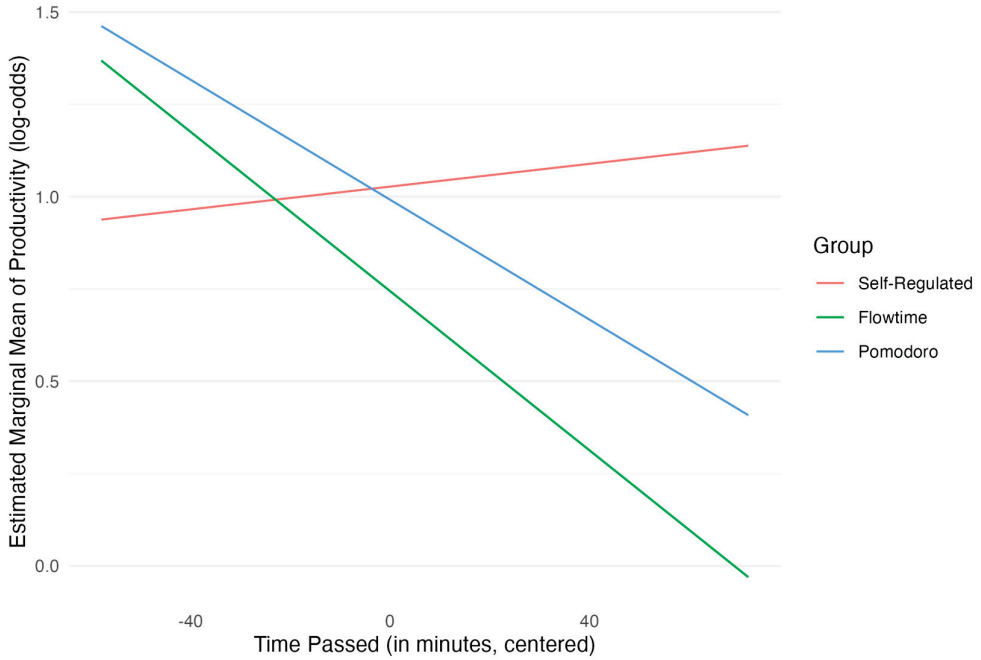


Figure 3. Estimated log-odds of productivity by group over time.

Overall, the hypothesis of direct group differences in productivity was not supported. There was a trend ($p = 0.051$) suggesting that productivity in the Flowtime group decreased more rapidly than in the self-regulated group as time since the last break increased. Furthermore, the results strongly underscore the importance of immediate history: lower prior fatigue, higher prior productivity, and higher prior motivation were all significant predictors of higher current productivity levels. As with the model for motivation, the findings in relation to the lagged productivity and motivation predictors should be handled with care due to potential separation issues, potentially driven by sparse data at the extremes of the outcome variable.

5.3. Fatigue

Hypothesis 1c predicted that participants in the self-regulated group would report higher fatigue compared with those in the Pomodoro and Flowtime groups. To test this, we modeled fatigue, assessed on a 4-point Likert scale, using a cumulative link mixed effects model with a loglog link function (please note that we are not representing odds ratios in this section, due to the nature of the loglog link function).

The final model included fixed effects for group membership, time (centered), duration of the last break (centered), orthogonal quadratic polynomial terms for lagged fatigue and lagged productivity, and the interaction between group membership and time. The model incorporated participant-level random intercepts and random slopes for the effect of centered break duration to account for within-subject variability and individual differences in the break duration effect. Table 4 summarizes the key fixed effects estimates from the model.

Table 4. Overview of coefficients for fatigue levels compared with self-regulated break taking.

Predictor	Estimate	Std. Error	*z* Value	*p*-Value
Pomodoro group	−0.11	0.19	−0.59	0.558
Flowtime group	−0.31	0.19	−1.57	0.117
Time (centered)	0.00	0.00	0.14	0.885
Last break duration (centered)	−0.04	0.03	−1.37	0.172
Productivity (linear)	−3.28	1.44	−2.28	0.023 *
Productivity (quadratic)	−0.47	1.33	−0.36	0.722
Fatigue (linear)	16.59	1.92	8.66	0.001 ***
Fatigue (quadratic)	3.19	1.59	2.01	0.045 *
Pomodoro × Time	0.01	0.00	2.25	0.025 *
Flowtime × Time	0.00	0.00	0.56	0.577

* $p < 0.05$. *** $p < 0.001$.

The model evaluation indicated that while the random effects structure was retained based on model-building steps, formal comparisons via AIC (714.2 vs. 709.6), BIC (776.4 vs. 760.1), and LRT ($\chi^2(3) = 1.38$, $p = 0.709$) suggested a simpler model without random effects (CLM) might offer comparable or slightly better fit based purely on these metrics. However, the CLMM was retained to explicitly account for the nested data structure inherent in the repeated measures design. The random intercept variance was relatively small ($\sigma^2_{\text{intercpt}} = 0.07$, $SD = 0.26$), as was the random slope variance for break duration ($\sigma^2_{\text{slope}} = 0.00$, $SD = 0.06$). A perfect negative correlation (−1.00) was estimated between the random intercept and slope, which might indicate boundary estimation issues, though the model converged successfully. Q-Q plots of random effects showed reasonable normality; although potential outliers were noted, they did not significantly alter results upon removal. The final model significantly outperformed a null model ($\chi^2(12) = 103.48$, $p < 0.001$). Tests of the proportional odds assumption (using a fixed-effects equivalent model) showed no significant violations overall, but the quadratic term for lagged fatigue could not be assessed directly in the nominal test output, and visual checks suggested potential non-proportionality at the lowest threshold, warranting caution in interpreting this specific term. The threshold coefficients that delineate the latent fatigue categories were estimated as 1|2: −1.55 ($SE = 0.18$), 2|3: −0.13 ($SE = 0.15$), and 3|4: 2.36 ($SE = 0.22$). The participant-level random effects estimates were intercept variance = 0.065 ($SD = 0.26$), break duration slope variance = 0.003 ($SD = 0.06$), correlation = −1.00.

Regarding Hypothesis 1c, the main effects of group were not statistically significant at the average level of centered time (the mid-point of the study session), indicating that baseline fatigue log-odds did not differ markedly between the self-regulated, Pomodoro ($\beta = -0.11$, $SE = 0.19$, $p = 0.558$), and Flowtime groups ($\beta = -0.30$, $SE = 0.19$, $p = 0.117$). The main effect for centered time since the last break was also not significant ($\beta = 0.00$, $SE = 0.00$, $p = 0.885$). However, a significant interaction emerged between the Pomodoro group and centered time since the last break ($\beta = 0.01$, $SE = 0.00$, $p = 0.025$). This indicates that, compared with the self-regulated group, the log-odds of reporting higher fatigue increased more steeply for participants in the Pomodoro group as more time elapsed since their last break. The interaction between the Flowtime group and time since the last break was not statistically significant ($\beta = 0.00$, $SE = 0.00$, $p = 0.577$).

Post hoc analyses confirmed the interaction pattern. Pairwise comparisons of the slopes showed that the rate of increase in fatigue log-odds over time was steeper for the Pomodoro group compared with the self-regulated group ($\Delta = 0.01$, $SE = 0.00$, $z = 2.25$, $p = 0.074$), although this comparison did not reach significance after Bonferroni adjustment. No other slope comparisons were significant after adjustment. Pairwise comparisons of estimated marginal means at specific time points revealed that at the third quartile

of time since the last break (ca. 120 min), the Pomodoro group had significantly higher estimated fatigue log-odds than the Flowtime group ($\Delta = 0.50$, $SE = 0.19$, $z = 2.60$, $p = 0.028$, Bonferroni-adjusted). Figure 4 shows the slopes for the three groups across time.

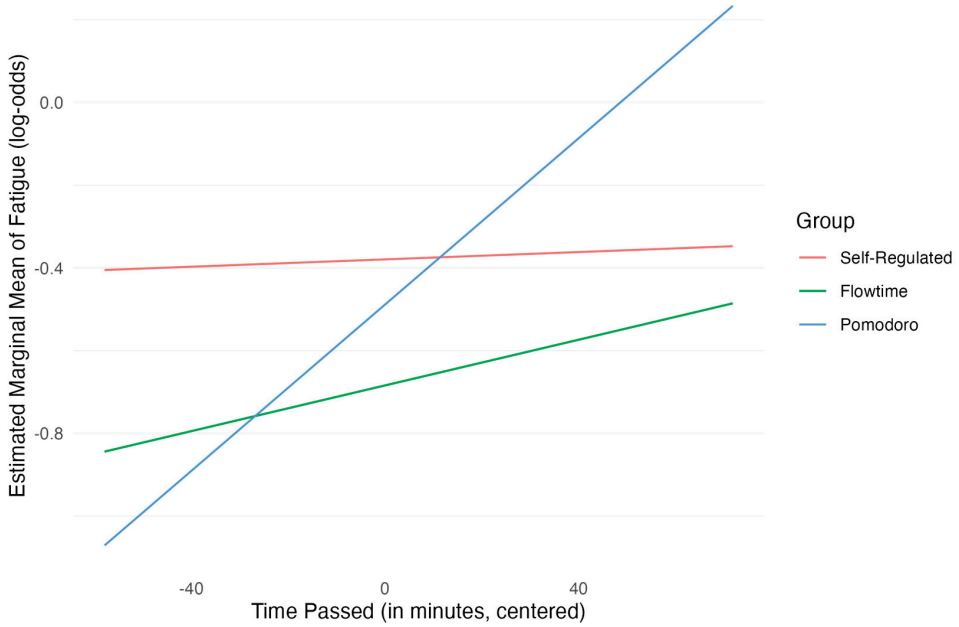


Figure 4. Estimated log-odds of fatigue by group over time.

Regarding other predictors, the centered duration of the last break did not show a significant fixed effect on fatigue ($\beta = -0.04$, $SE = 0.03$, $p = 0.172$), although its effect was kept as a random slope based on improved model fit (as indicated by AIC and BIC). Lagged fatigue showed a strong positive association with current fatigue, significant for both the linear ($\beta = 16.59$, $SE = 1.92$, $p < 0.001$) and quadratic terms ($\beta = 3.19$, $SE = 1.59$, $p = 0.045$), indicating a positive effect of prior fatigue (linear term) that is particularly strong at extreme levels of fatigue (quadratic term). This might be driven by a potential floor effect in the case of the lower extreme (e.g., an increased likelihood of an increase in fatigue when one is experiencing low fatigue at $t - 1$) and an especially negative impact of very high fatigue on fatigue at subsequent time points. Lagged productivity showed a significant negative linear association with fatigue ($\beta = -3.28$, $SE = 1.44$, $p = 0.022$), suggesting higher prior productivity predicted lower current fatigue; the quadratic term was not significant ($p = 0.722$).

Overall, these results do not support the hypothesis of consistently lower fatigue in the structured break groups (Pomodoro, Flowtime) compared with the self-regulated group based on main effects. Instead, fatigue levels were strongly predicted by prior fatigue and prior productivity. Crucially, a significant interaction revealed that fatigue increased more rapidly for the Pomodoro group compared with the self-regulated group as time elapsed since the last break.

5.4. Task Completion

Hypothesis 2 predicted that the Pomodoro and Flowtime conditions would show higher levels of task completion compared with the self-regulated break condition. To test this hypothesis, we employed an Analysis of Covariance (ANCOVA). Task completion (the percent of overall completed tasks) served as the outcome variable. The predictors included group, task amount (the number of planned tasks), total duration spent studying, and the aggregated measures of motivation and productivity. Table 5 presents the ANCOVA table. Additional model diagnostics, including correlation matrix, squared semi-partial correlations, and several diagnostic plots for this model, can be found in Appendix C.

Table 5. Overview of ANCOVA model for task completion levels compared with self-regulated break taking.

Predictor	SS	DF	*F*-Value	*p*-Value
Group	105.61	2	0.16	0.854
Task Amount	35.41	1	0.11	0.746
Total Study Time	739.97	1	2.21	0.141
Motivation (Avg.)	669.25	1	2.00	0.161
Productivity (Avg.)	1774.56	1	5.30	0.024 *
Residuals	29,148.23	87	NA	NA

* $p < 0.05$.

The overall ANCOVA model explained approximately 8.7% of the variance in task completion (Multiple $R^2 = 0.09$, Adjusted $R^2 = 0.02$), although the *F*-test ($F(6, 87) = 1.38$, $p = 0.231$) indicated that the model did not reach overall statistical significance.

To assess the unique contribution of each predictor after accounting for others, we examined Type II Sum of Squares from the ANCOVA (see Table 5). This analysis revealed that the primary predictor of interest, group, did not have a statistically significant effect on task completion ($F(2, 87) = 0.16$, $p = 0.854$). Among the predictors, productivity emerged as the only statistically significant predictor of task completion ($F(1, 87) = 5.30$, $p = 0.024$). The estimated coefficient of the productivity term ($\beta = 12.31$) suggests that holding all other variables constant, a 1-unit increase in the aggregated productivity measure is associated with an increase of approximately 12.31 percentage points in the overall task completion rate.

Therefore, these findings do not support Hypothesis 2. While break-taking techniques per se did not lead to significant group differences in task completion, higher average productivity appears to be an important factor in predicting the percentage of completed tasks. The overall insignificance of the ANCOVA model and the large residual term indicate that these conclusions should be treated with great care.

5.5. Flow State

Hypothesis 3 proposed that the flow state, measured by the Flow State Scale (FSS) total score, would be greater in the self-regulated break and Flowtime conditions than in the Pomodoro condition. To test this, we conducted an ANCOVA with the overall flow score as the outcome. The predictors included group, the number of planned tasks, the total duration spent studying, and the aggregated measures of motivation and productivity. Table 6 presents the ANCOVA table. Additional model diagnostics, including correlation matrix, squared semi-partial correlations, and several diagnostic plots for this model, can be found in Appendix C.

Table 6. ANCOVA table for flow state compared with Pomodoro.

Predictor	SS	DF	*F*-Value	*p*-Value
Group	0.23	2	0.26	0.774
Task Amount	3.34	1	7.40	0.008 **
Total Study Time	0.50	1	1.10	0.297
Motivation (Avg.)	0.28	1	0.62	0.433
Productivity (Avg.)	3.65	1	8.09	0.006 **
Residuals	39.23	87	NA	NA

** $p < 0.01$.

The ANCOVA model demonstrated a Multiple R^2 of 0.31 (Adjusted R-squared = 0.26). An overall *F*-test indicated that the model was statistically significant ($F(6, 87) = 6.41$, $p < 0.001$).

Examining the unique contribution of each predictor using Type II Sum of Squares (see Table 6) revealed that the primary predictor of interest, group, did not have a statistically significant overall effect on flow scores ($F(2, 87) = 0.26$, $p = 0.774$) after controlling for the covariates. However, both task amount ($F(1, 87) = 7.40$, $p = 0.008$) and productivity ($F(1, 87) = 8.09$, $p = 0.006$) emerged as significant predictors. The respective estimated coefficients suggest that a 1-unit increase in task amount was associated with a 0.12 increase in the flow score, while a 1-unit increase in productivity was linked to a 0.56 increase in flow. In contrast, neither total time spent studying ($p = 0.297$) nor motivation ($p = 0.433$) contributed significantly to explaining the variance in flow scores.

Therefore, these findings do not support Hypothesis 3. While the hypothesized group differences in flow were not found, higher numbers of planned tasks and greater aggregated productivity significantly predicted increased flow levels. The overall model accounted for approximately 31% of the variance in flow, indicating a moderate explanatory power for these factors in relation to the flow experience, but the specific break-taking technique assigned did not significantly influence reported flow.

6. Discussion

This study examined the effect on motivation, productivity, fatigue, task completion, and flow of taking self-regulated, Pomodoro, or Flowtime breaks during a self-regulated learning session. Furthermore, we examined how these different break-taking techniques influence the trajectory of subjective study experiences over time.

Hypothesis 1, which stated that the Pomodoro and Flowtime conditions would show higher levels of motivation and productivity and lower levels of fatigue compared with the self-regulated condition, was not supported by our findings. No significant difference in motivation, productivity, or fatigue was found between the break-taking techniques.

However, it was found that the motivation of students who used the Pomodoro and Flowtime techniques declined faster compared with the self-regulated group. This might be explained by the fact that students in these conditions were not allowed to regulate the duration of the break by themselves. This could mean that the breaks are of suboptimal length for some students, which may lead to frustration. Consequently, this may in turn lead to more demotivation (Deci & Ryan, 2000).

Furthermore, it has been found that students who took breaks according to the Pomodoro group showed a faster increase in fatigue levels. This could again be the case because students were not allowed to plan the duration of the breaks in this condition, which may have led to a suboptimal duration of breaks in which students were not able to restore their energy levels sufficiently (Lim & Kwok, 2016; Trougakos & Hideg, 2009).

Additionally, it was also found that the motivation of students is predicted by prior fatigue, productivity, and motivational levels. Lower fatigue, higher prior productivity, and higher motivation also significantly predicted higher productivity levels. These subjective experiences thus have an influence on each other, which means that it is important for students to restore all of these resources in order to be able to study effectively (Li et al., 2020; Lyubykh et al., 2022; Phan & Beck, 2023).

Hypothesis 2, which stated that the Pomodoro and Flowtime conditions would show higher levels of task completion compared with the self-regulated condition, was also not supported by our data. No significant differences in task completion levels between groups were found. However, it seems like productivity is an important predictor of task completion. Students who experienced higher levels of productivity completed a higher percentage of their planned tasks.

The current study findings suggest a minimal impact of different break-taking techniques on subjective study experiences and task completion. Considerable variability

among participants within each break condition was observed, likely influenced by several factors. First of all, break effectiveness depends highly on the nature of activities and experiences during the break (Kim et al., 2016; Melguizo-Ibáñez et al., 2024; Sammonds et al., 2017). Prior research suggests that active breaks, such as engaging in brief physical activities, are found to enhance attention and task performance (Melguizo-Ibáñez et al., 2024; Sammonds et al., 2017). Furthermore, relaxation and socialization during break time can also enhance break effectiveness (Ginns et al., 2023; Kim et al., 2016). In the current study, however, there was no control over the content of the breaks, and no data was gathered on what participants did during the breaks. The variability in the effectiveness of the breaks may therefore be partly explained by a variability in break activities and experiences. The effects of different break activities across these break-taking techniques may be investigated by future research.

In addition, individual differences may also influence the effectiveness of a certain break-taking technique (Fritz et al., 2013; Trougakos & Hideg, 2009). Research shows that people's level of respite self-efficacy (confidence in recovery during a break) has been linked to break effectiveness in occupational settings. Lower levels of respite self-efficacy are associated with poorer detachment from tasks and reduced break effectiveness (Fritz et al., 2013). Moreover, personality traits may influence personal resource depletion and therefore the need for breaks (Fostervold & Watten, 2024; Trougakos & Hideg, 2009). Research suggests that people high in emotional stability experience a reduced need for recovery after work (Fostervold & Watten, 2024). Furthermore, extroverts may require fewer and/or shorter breaks than introverts when performing demanding tasks, as they experience lower personal resource depletion (Trougakos & Hideg, 2009). These findings suggest that the effectiveness of different break-taking techniques may depend on or interact with individual characteristics. This might indicate that different break-taking strategies may work for different people, which may be investigated in future research.

Lastly, the small differences between the break-taking techniques may be explained by the fact that flow was experienced equally in different conditions. The flow state was thus not more disrupted in the Pomodoro condition compared with other break-taking conditions. In the entire sample of this study, flow was not experienced to a great extent. On a scale from 0 to 7, the average level of flow was 4.06 in this sample. A possible explanation for this is that it is important to create a balance between challenge and skill in order to reach a flow state (Willekes, 2014; Zielke et al., 2023). This is because when a task is too easy people will feel bored, but if a task is too difficult people can become stressed, anxious, or feel overwhelmed. In both these cases, a flow state will not be reached (Fong et al., 2015). It could be the case that this balance is not optimal when engaging in study tasks. University assignments are normally more challenging than the current skill level of students (Guadagnoli et al., 2012). This is so that students' skill levels will increase over time. This could explain why, in this sample, the flow state experienced was low in all conditions. The Flowtime technique therefore may not have been found to be more effective than the Pomodoro technique.

6.1. Limitations and Future Research

This study comes with several limitations. Due to the authentic learning setting of this study, students could choose when and where to study, which inherently limited experimental control. Students could not be monitored, and therefore, this study setup required the assumption that students would follow the instructions given to them. Another limitation of the authentic learning setting is that all students can choose their own study tasks. This way, not every student worked on a study task of the same difficulty, which might have influenced the dependent variables. Even though we randomized the allocation

of students to conditions, the sample size might have been too small to distribute the potential effects of differences in task difficulty across conditions. Future studies might consider keeping the study tasks constant for all students. This will overcome these drawbacks of the authentic study setting, which can make it easier to compare break-taking techniques between participants.

To maintain the authenticity of the study sessions, interruptions were minimized as much as possible. Because of this, single-item measures were used to assess motivation, productivity, and fatigue. Even though these measures may not have been optimal to measure the complexity of these constructs, more extensive measures were avoided to prevent excessive disruption during the students' study sessions. By using single items, the interruption between study sessions and load for students was also minimized. Nevertheless, students who took more breaks were required to complete the self-evaluations more often, which may have influenced their study and break behavior (Pintrich, 2000). However, if this had significantly affected outcomes, it could be expected that clearer differences between the conditions would have been observed, which was not the case in this study.

Additionally, participants were informed before the study that the study was about break taking. This could have lowered the validity of the study and could have made the students more aware of their break taking (Rothstein & Shoben, 2013). However, participants were not informed about the different conditions and the goal of the study until after the study was finished, which allowed for minimal interference.

The questionnaire was programmed in such a way that the questionnaire would continue automatically after 2 h of study. However, this function did not work for all participants, probably due to different operating systems on students' computers and the tasks students performed on their laptops during the study session. Some students therefore needed to manually click a link to the last questions of the survey. This may have led to some limitations, such as students needing to check the time themselves, which could have made them more aware of the time they study.

Furthermore, results regarding feelings of productivity should be treated with care because the proportional hazards assumption might be violated for low levels of productivity. However, given that relatively few participants scored low on productivity, the findings related to productivity levels still hold at higher productivity levels.

Lastly, the current sample introduces some limitations to this study. The final sample of the study, after the removal of outliers, consisted of 94 participants. Because of the shift in analysis approach, this may underpower the analysis, especially for the CLMM analysis. Future research should aim for a bigger sample size to ensure more stable model estimates. Furthermore, the sample of this study is very homogeneous, with approximately 93 percent women, all of whom were psychology bachelor students. Since women are also highly overrepresented in psychology studies, this sample represents the underlying population well (Centraal Bureau voor de Statistiek, 2023). However, this sample cannot be considered representative of the general student population. We therefore refrain from making generalized statements beyond the specific context of this sample. Future studies should focus on measuring more heterogenous samples so that they can investigate whether the same results are found in other genders and study areas.

Future studies are necessary to investigate the effectiveness of these break-taking techniques further. Future studies can focus on longer study times. By doing this, it can be investigated whether these techniques are efficient for longer sessions. This way it can be investigated whether the development in time on levels of motivation, productivity, and fatigue will continue in a similar manner or not.

Additionally, future studies might focus on mental effort and personal preferences as predictors for the effectiveness of different break-taking techniques to investigate if these are significant contributors to the effectiveness of different techniques.

6.2. Conclusions

To conclude, the current findings suggest that for students using Pomodoro and Flowtime techniques, motivation levels decrease faster compared with students who use self-regulated breaks. Similarly, students who used the Pomodoro technique showed a steeper increase in fatigue levels. However, despite these significant differences in slopes of motivation and fatigue over time, these differences did not result in significant differences between the break-taking techniques in the overall means of motivation and fatigue levels. This might indicate that the 2 h study duration was not long enough to find differences among these break techniques. On all other variables measured, no outstanding differences were found in the effectiveness of self-regulated, Pomodoro, and Flowtime breaks. This means that the results of the current study suggest that no one of these break-taking techniques should be more highly recommended over the others. Future research should be performed to investigate the differences between these techniques further by including predictors like personality and mental effort and by examining longer study periods.

Author Contributions: Conceptualization, E.J.C.S., N.W. and A.d.B.; methodology, E.J.C.S. and N.W.; software, E.J.C.S.; validation, E.J.C.S. and N.W.; formal analysis, N.W.; investigation, E.J.C.S.; resources, E.J.C.S. and N.W.; data curation, E.J.C.S. and N.W.; writing—original draft preparation, E.J.C.S. and N.W.; writing—review and editing, N.W. and A.d.B.; visualization, E.J.C.S.; supervision, N.W. and A.d.B.; project administration, E.J.C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Review Committee Psychology and Neuroscience (ERCPN), Maastricht University (reference code: ERCPN-274_120_11_2023, date of approval: 6 December 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets presented in this article are not readily available because the data was collected as part of a bachelor thesis, and therefore, the ethical approval did not include availability of the data. Requests to access the datasets should be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Instructions for Participants About Break-Taking Technique

Participants were randomly assigned to one of the following three blocks in which they were given the instruction to study for two hours and the information about the break-taking technique they should use.

Appendix A.1. Self-Regulated Break-Taking Instructions

In the following 2 h you are going to do self-study. Please work on the tasks you have described in the previous question. In these 2 h you are asked to study for as long as you can focus. If you feel like you need a break you can press the arrow. Take a break as long as you need. If you feel like you are ready to start studying again, please press the arrow again. After this you will be asked to fill in some questions and then, after you pressed the arrow again you will start to study again. After two hours the questionnaire automatically moves to the last part of this questionnaire.

If you heard a sound when this page was shown, the same sound will play if the two hours of self-study are over. If you did not hear the sound, please pay attention yourself when the page changes to the last part. Besides that, if you open another tab, the sound will also not play automatically, in that case please also pay attention yourself when the page changes.

Good luck with studying!

Appendix A.2. Pomodoro Break Instructions

In the following 2 h you are going to do self-study. Please work on the tasks you have described in the previous question. In these 2 h you will study in blocks of 25 min with a 5 min break in between. This questionnaire will show you what to do at what time. The questionnaire will automatically transfer to the break after 25 min. Please take a break from your studies when it is said. After the break you will be asked to fill in some questions and press the arrow again. After this you will be instructed to start studying again. After two hours the questionnaire automatically moves to the last part of this questionnaire.

If you heard a sound when this page was shown, the same sound will play whenever you need to change your activity (so if you need to start studying after a break or if you need to take a break). If you did not hear the sound, please pay attention yourself when the page changes to break time or study time. Besides that, if you open another tab, the sound will also not play automatically, in that case please also pay attention yourself when the page changes.

Good luck with studying!

Appendix A.3. Flowtime Break Instructions

In the following 2 h you are going to do self-study. Please work on the tasks you have described in the previous question. In these 2 h you are asked to study for as long as you can focus. If you feel like you need a break you can press the arrow. The break you will get is predetermined; the page will automatically transfer to another page after your break time is over. After the break you will be asked to fill in some questions and then you will start to study again. After two hours the questionnaire automatically moves to the last part of this questionnaire.

If you heard a sound when this page was shown, the same sound will play whenever you need to change your activity (so if you need to start studying after a break). If you did not hear the sound, please pay attention yourself when the page changes to study time. Besides that, if you open another tab, the sound will also not play automatically, in that case please also pay attention yourself when the page changes.

Good luck with studying!

Appendix B. Full Questionnaire Measures

Appendix B.1. Assessment Motivation, Productivity, and Fatigue Levels

Motivation, productivity, and fatigue levels were assessed using the format shown in Figure A1. These questions were asked before the start of the study period, every time students took a break, and after the study period. The participants were asked to "Fill in the following statements about your state of mind at this moment".

	Not at all	Very little	Somewhat	To a great extent
How fatigued do you feel right now?	O	O	O	O
How motivated do you feel right now?	O	O	O	O
How productive do you feel right now?	O	O	O	O

Figure A1. Assessment of motivation, productivity, and fatigue levels.

Appendix B.2. Assessment Task Completion

Appendix B.2.1. Assessment Task Completion Before Study Time

Task completion was assessed in the format shown in Figure A2. The participants were asked, "What tasks are you planning to do in the following hours of self-study? Please be specific. If you have planned less than 8 tasks only fill in the amount of tasks and a description of the task you have planned (you do not have to fill in all the 8 lines)".

Task 1	
Task 2	
Task 3	
Task 4	
Task 5	
Task 6	
Task 7	
Task 8	

Figure A2. Assessment of task completion before study time.

Appendix B.2.2. Assessment Task Completion After Study Period

Task completion after the study session was assessed in the format shown in Figure A3. The participants were asked to "Fill in to which extent you completed the tasks you planned to do in this self-study time from 0% (did not do anything for this task) until 100% (this task is completely done)".

The tasks that were filled in before the study time were forwarded to this question. The names "Task 1" to "Task 8" were therefore replaced with the planned tasks of the participant.

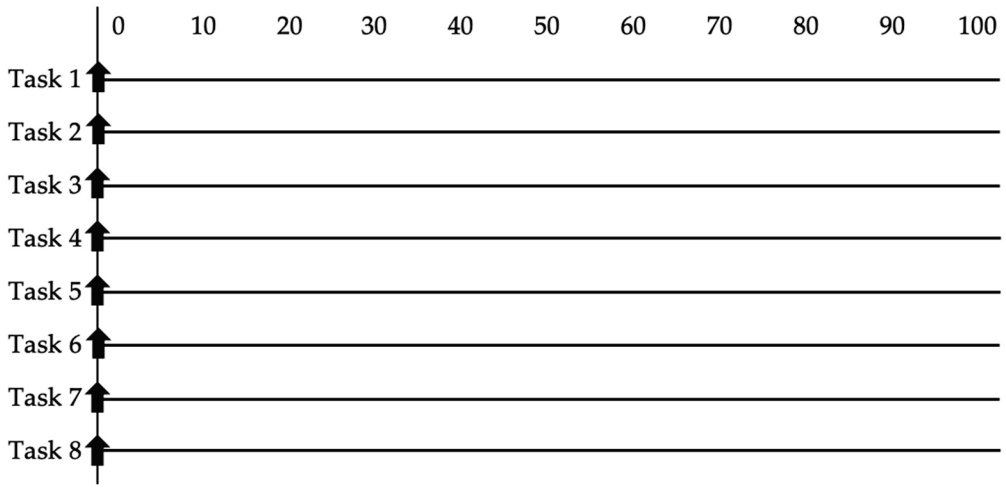


Figure A3. Assessment of task completion after the study period.

Appendix B.3. Assessment Flow State: Flow Short Scale (FSS)

Flow state was assessed in the format shown in Figure A4. The participants were asked to "Fill in the following statements about your feelings in the last 2 h during study time".

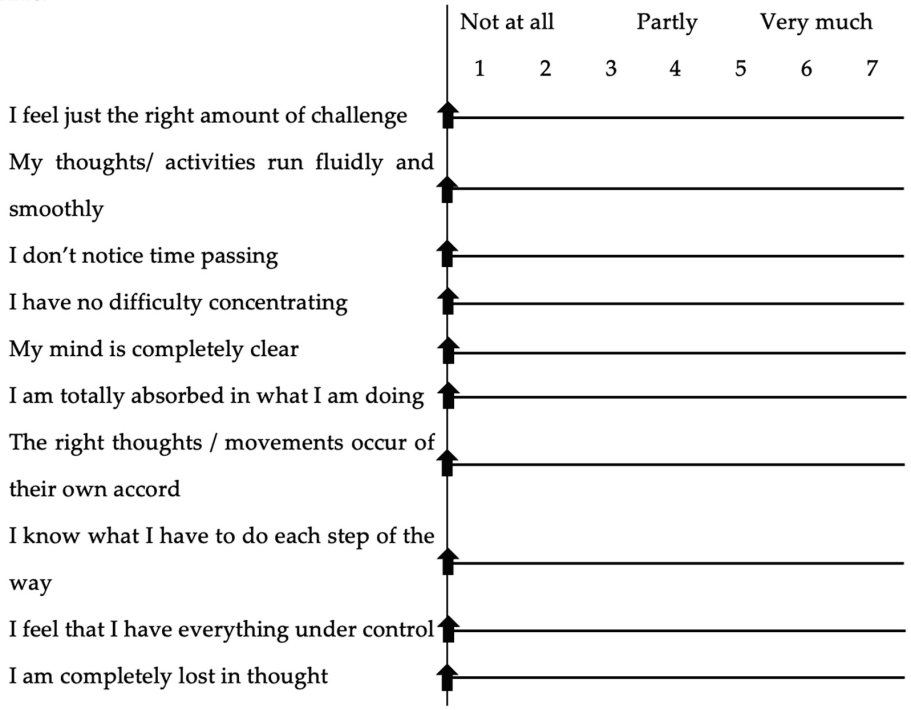


Figure A4. Assessment of flow state.

Appendix C. Model Diagnostics

Appendix C.1. General Diagnostics

Table A1. Means, standard deviations, and correlations with confidence intervals.

Variable	*M*	*SD*	1	2	3	4	5	6
1. Task Completion (%)	68.84	18.53						
2. Flow (FSS)	4.09	0.78	0.15 [−0.06, 0.34]					
3. Task Amount	3.48	1.64	−0.07 [−0.27, 0.14]	0.17 [−0.03, 0.36]				
4. Study Duration	1.70	0.15	0.14 [−0.07, 0.33]	0.08 [−0.13, 0.28]	−0.09 [−0.29, 0.11]			
5. Motivation (Avg.)	2.72	0.56	0.09 [−0.11, 0.29]	0.43 ** [0.25, 0.58]	−0.12 [−0.31, 0.09]	0.08 [−0.12, 0.28]		
6. Productivity (Avg.)	2.87	0.58	0.21 * [0.01, 0.40]	0.49 ** [0.32, 0.63]	−0.10 [−0.30, 0.10]	0.01 [−0.20, 0.21]	0.79 ** [0.70, 0.85]	

M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation. * indicates $p < 0.05$. ** indicates $p < 0.01$.

Table A2. VIF for predictors of task completion and flow.

Variable	GVIF	Df	$GVIF^{\frac{1}{2Df}}$
Group	1.09	2	1.02
Task Amount	1.06	1	1.03
Study Duration	1.03	1	1.03
Motivation (Avg.)	2.71	1	1.65
Productivity (Avg.)	2.68	1	1.64

Appendix C.2. Diagnostics for Task Completion

Table A3. Squared semi-partial correlations for task completion.

Variable	sr^2	Confidence Interval
Group	0.003	[0.00, 0.03]
Task Amount	0.001	[0.00, 0.01]
Study Duration	0.020	[0.00, 0.08]
Motivation (Avg.)	0.020	[0.00, 0.08]
Productivity (Avg.)	0.060	[0.00, 0.14]

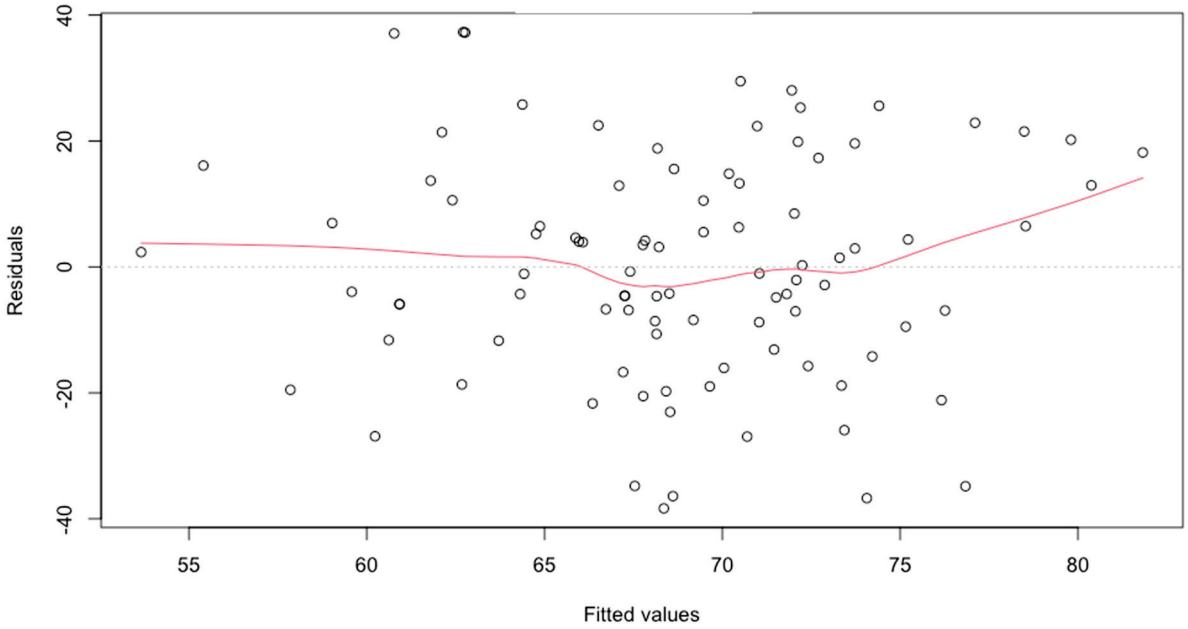


Figure A5. Residual plot for task completion model.

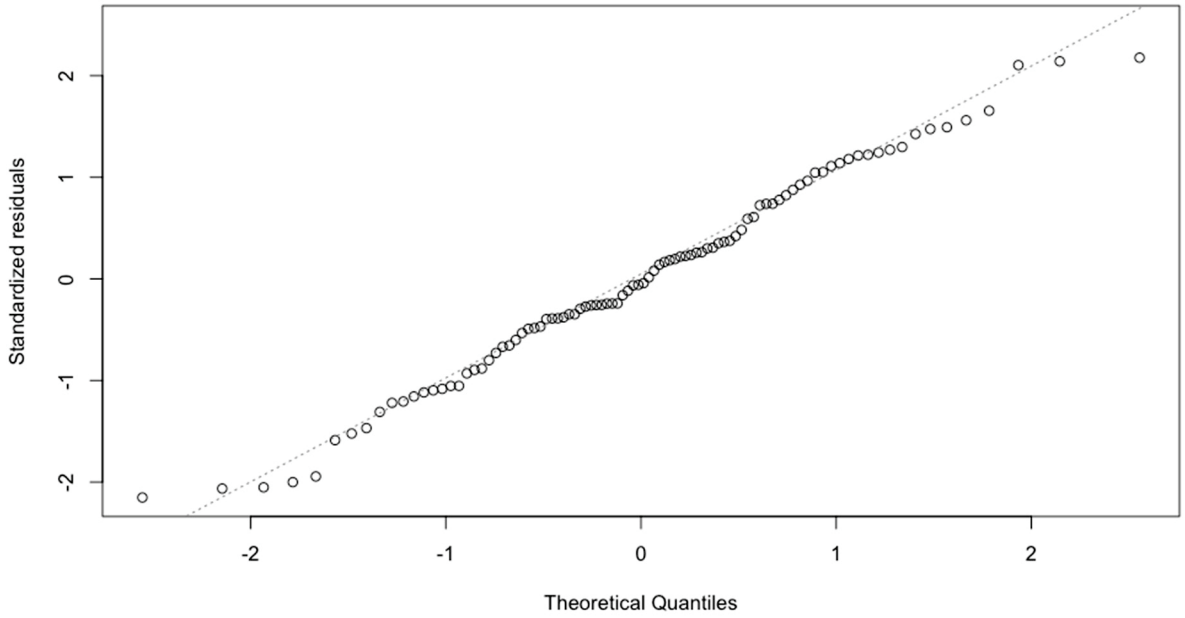


Figure A6. Normal Q-Q plot for task completion model.

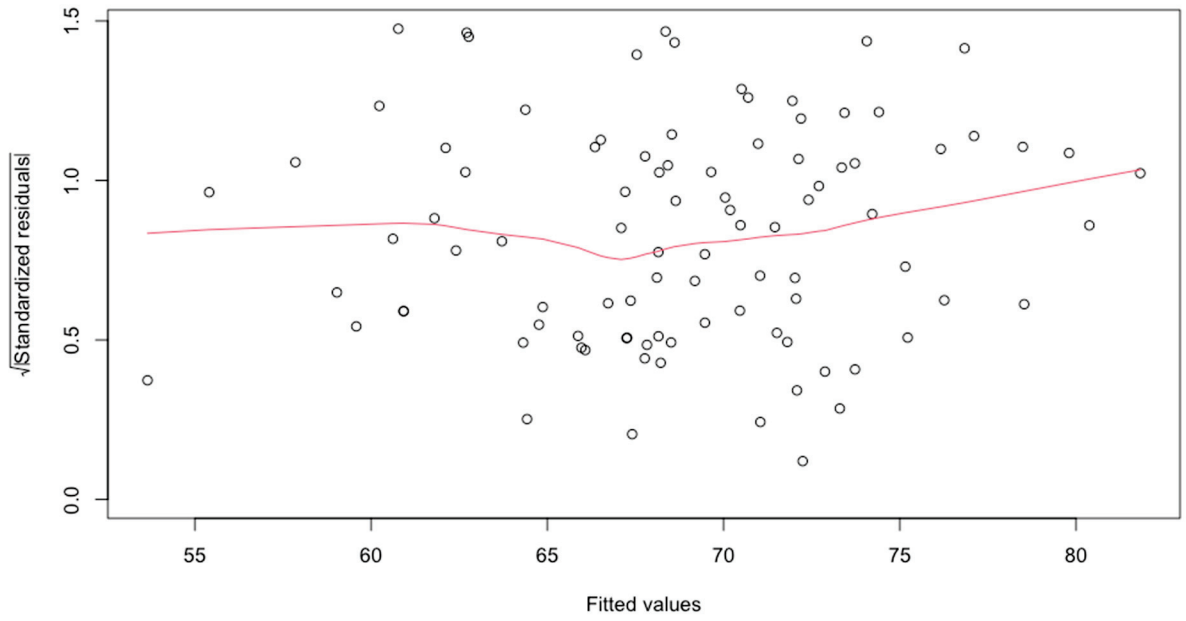


Figure A7. Scale-location plot for task completion model.

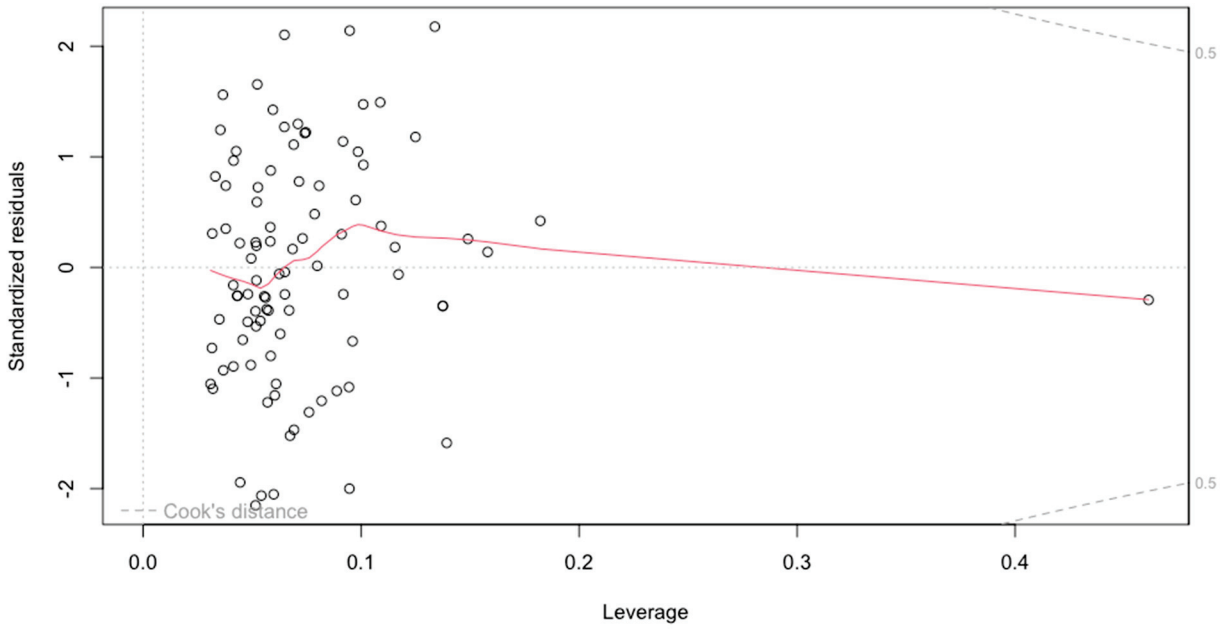


Figure A8. Residual vs. leverage plot for task completion model.

Appendix C.3. Diagnostics for Flow State

Table A4. Squared semi-partial correlations for flow state.

Variable	sr^2	Confidence Interval
Group	0.004	[0.00, 0.03]
Task Amount	0.060	[0.00, 0.14]
Study Duration	0.009	[0.00, 0.04]
Motivation (Avg.)	0.005	[0.00, 0.03]
Productivity (Avg.)	0.060	[0.00, 0.15]

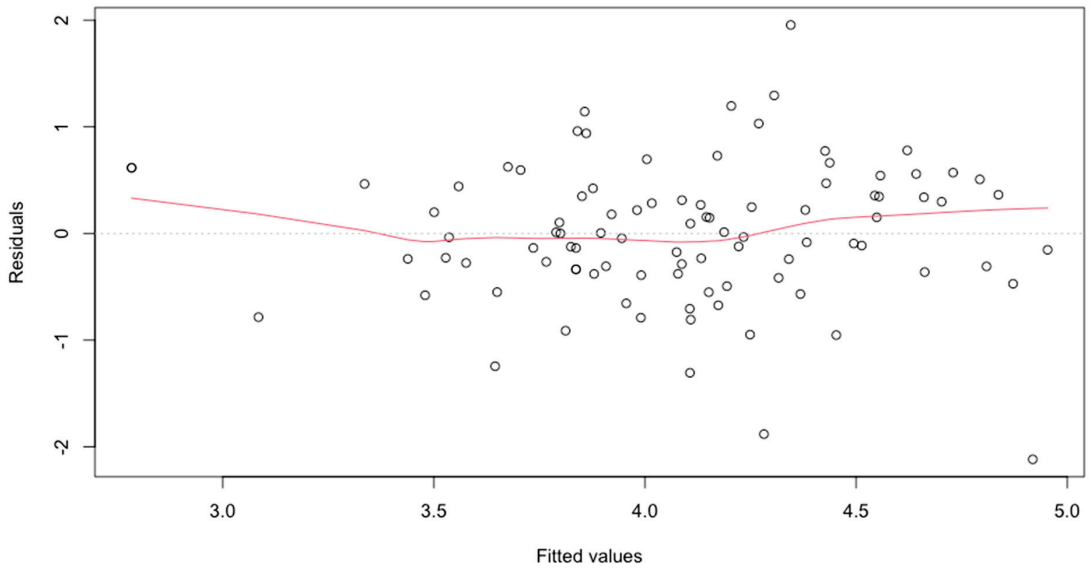


Figure A9. Residual plot for flow state model.

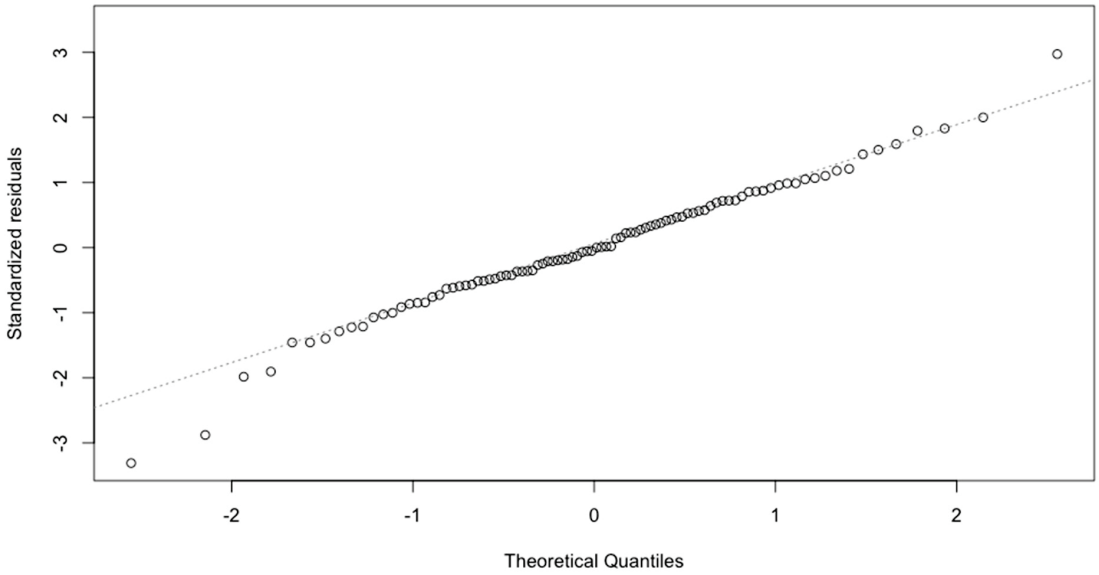


Figure A10. Normal Q-Q plot for flow state model.

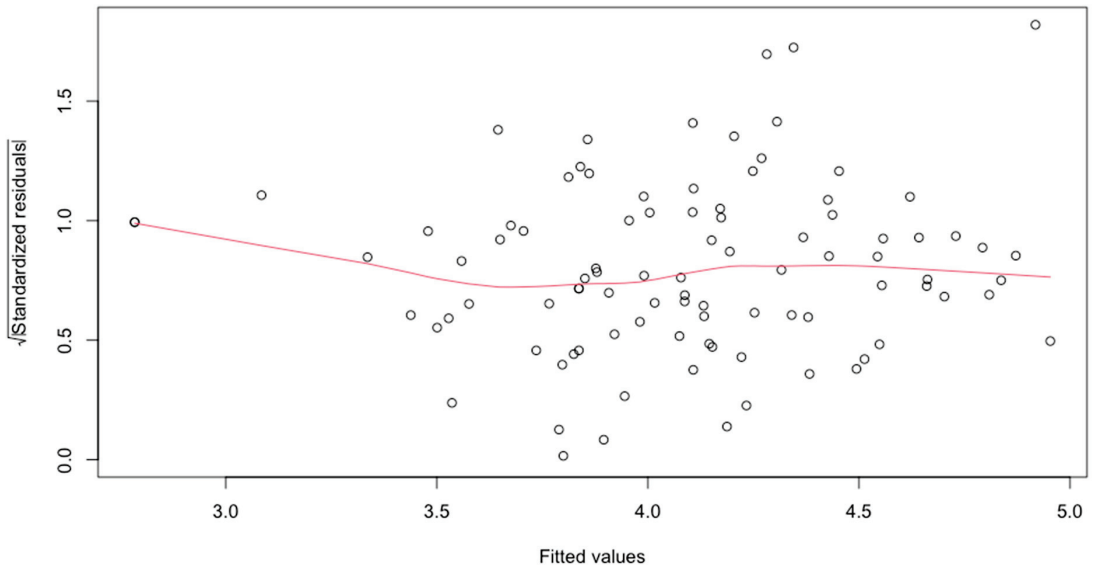


Figure A11. Scale-location plot for flow state model.

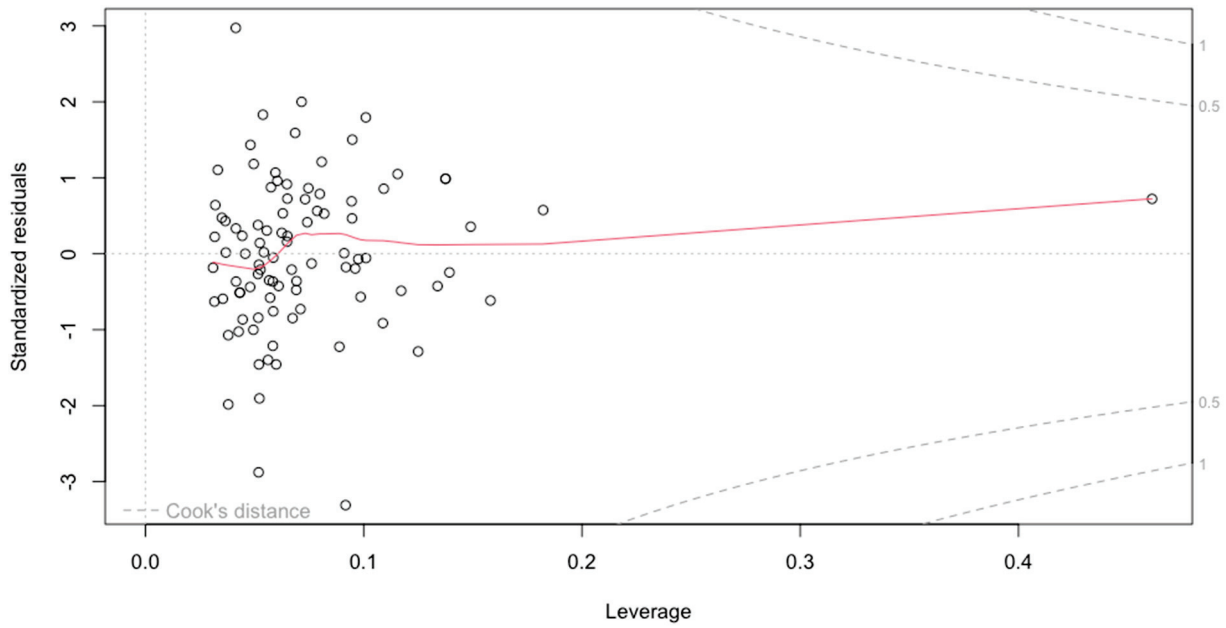


Figure A12. Residual vs. leverage plot for flow state model.

References

Abad, Z. S. H., Karras, O., Schneider, K., Barker, K., & Bauer, M. (2018). Task interruption in software development projects: What makes some interruptions more disruptive than others? *arXiv*, arXiv:1805.05508.

Almalki, K., Alharbi, O., Al-Ahmadi, W., & Aljohani, M. (2020). Anti-procrastination online tool for graduate students based on the Pomodoro technique. In *Learning and collaboration technologies* (Vol. 12206, pp. 133–144). Springer. [CrossRef]

Biwer, F., de Bruin, A., & Persky, A. (2023a). Study smart—Impact of a learning strategy training on students' study behavior and academic performance. *Advances in Health Sciences Education*, *28*(1), 147–167. [CrossRef] [PubMed]

Biwer, F., Wiradhany, W., Oude Egbrink, M. G. A., & de Bruin, A. B. H. (2023b). Understanding effort regulation: Comparing 'Pomodoro' breaks and self-regulated breaks. *British Journal of Educational Psychology*, *93*(S2), 353–367. [CrossRef]

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, *9*(4), 475–479. [CrossRef]

Centraal Bureau voor de Statistiek. (2023, March 7). *Al 23 jaar op rij meer vrouwen dan mannen in hoger onderwijs*. Centraal Bureau Voor De Statistiek. Available online: https://www.cbs.nl/nl-nl/nieuws/2023/10/al-23-jaar-op-rij-meer-vrouwen-dan-mannen-in-hoger-onderwijs (accessed on 7 May 2025).

Christensen, R. H. B. (2023). *Regression models for ordinal data (v. 12-4.1)*. Available online: https://cran.r-project.org/web/packages/ordinal/index.html (accessed on 20 January 2025).

Csikszentmihalyi, M. (1975). Play and intrinsic rewards. *Journal of Humanistic Psychology*, *15*(41), 41–63. [CrossRef]

de Bruin, A. B. H., Biwer, F., Hui, L., Onan, E., David, L., & Wiradhany, W. (2023). Worth the effort: The start and stick to desirable difficulties (S2D2) framework. *Educational Psychology Review*, *35*(2), 41. [CrossRef]

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268. [CrossRef]

Dizon, R. J., Ermitanio, H. D., Estevez, D. M., Ferrer, J., Flores, S. J., Fontanilla, K. M., Frias, A., Galang, E., Guei, N. F., & Sugay, J. (2021). The effects of Pomodoro technique on academic-related tasks, procrastination behavior, and academic motivation among college students in a mixed online learning environment. *Globus Journal of Progressive Education*, *11*(2), 58–63. [CrossRef]

Dora, J., van Hooff, M. L. M., Geurts, S. A. E., Kompier, M. A. J., & Bijleveld, E. (2022). The effect of opportunity costs on mental fatigue in labor/leisure trade-offs. *Journal of Experimental Psychology: General*, *151*(3), 695–710. [CrossRef]

Fong, C. J., Zaleski, D. J., & Leach, J. K. (2015). The challenge–skill balance and antecedents of flow: A meta-analytic investigation. *The Journal of Positive Psychology*, *10*(5), 425–446. [CrossRef]

Fostervold, K. I., & Watten, R. G. (2024). Put your feet up: The impact of personality traits, job pressure, and social support on the need for recovery after work. *Current Psychology*, *43*, 8485–8501. [CrossRef] [PubMed]

Fritz, C., Ellis, A. M., Demsky, C. A., Lin, B. C., & Guros, F. (2013). Embracing work breaks: Recovering from work stress. *Organizational Dynamics*, *42*(4), 274–280. [CrossRef]

Ginns, P., Muscat, K., & Naylor, R. (2023). Rest breaks aid directed attention and learning. *The Educational and Developmental Psychologist*, *40*(2), 141–150. [CrossRef]

Goetz, T., Nett, U. E., & Hall, N. C. (2013). Self-regulated Learning. In C. H. Hall, & T. Goutz (Eds.), *Emotion, motivation, and self-regulation: A handbook for teachers* (pp. 123–166). Emerald Group Publishing.

Grahek, I., Frömer, R., Prater Fahey, M., & Shenhav, A. (2022). Learning when effort matters: Neural dynamics underlying updating and adaptation to changes in performance efficacy. *Cerebral Cortex*, *33*(5), 2395–2411. [CrossRef]

Guadagnoli, M., Morin, M., & Dubrowski, A. (2012). The application of the challenge point framework in medical education. *Medical Education*, *46*(5), 447–453. [CrossRef]

Katidioti, I., Borst, J. P., van Vugt, M. K., & Taatgen, N. A. (2016). Interrupt me: External interruptions are less disruptive than self-interruptions. *Computers in Human Behavior*, *63*, 906–915. [CrossRef]

Kim, S., Park, Y., & Niu, Q. (2016). Micro-break activities at work to recover from daily work demands. *Journal of Organizational Behavior*, *38*(1), 28–44. [CrossRef]

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6), 661–679. [CrossRef]

Lau, J. (2023, July 6). *How to use the flowtime technique to boost your productivity*. Zapier. Available online: https://zapier.com/blog/flowtime-technique/ (accessed on 29 January 2025).

Lee, J. Y. (2022). *The medical pause in simulation training* [Doctoral thesis, Maastricht University]. [CrossRef]

Lee, J. Y., Szulewski, A., Young, J. Q., Donkers, J., Jarodzka, H., & van Merriënboer, J. J. G. (2021). The medical pause: Importance, processes and training. *Medical Education*, *55*(10), 1152–1160. [CrossRef]

Li, K., Xu, S., & Fu, H. (2020). Work-break scheduling with real-time fatigue effect and recovery. *International Journal of Production Research*, *58*(3), 689–702. [CrossRef]

Lim, J., & Kwok, K. (2016). The effects of varying break length on attention and time on task. *Human Factors*, *58*(3), 472–481. [CrossRef]

Lyubykh, Z., Gulseren, D., Premji, Z., Wingate, T. G., Deng, C., Bélanger, L. J., & Turner, N. (2022). Role of work breaks in well-being and performance: A systematic review and future research agenda. *Journal of Occupational Health Psychology*, *27*(5), 470–487. [CrossRef] [PubMed]

Melguizo-Ibáñez, E., Zurita-Ortega, F., González-Valero, G., Puertas-Molero, P., Tadeu, P., Ubago-Jiménez, J. L., & Alonso-Vargas, J. M. (2024). Active break as a tool for improving attention in the educational context. A systematic review and meta-analysis. *Revista de Psicodidáctica (English Ed.)*, *29*(2), 147–157. [CrossRef]

Oliveira dos Santos, W., Bittencourt, I. I., Isotani, S., Dermeval, D., Brandão Marques, L., & Frango Silveira, I. (2018). Flow theory to promote learning in educational systems: Is it really relevant? *Revista Brasileira de Informática Na Educação*, *26*(02), 29. [CrossRef]

Phan, V., & Beck, J. W. (2023). Why do people (not) take breaks? An investigation of individuals' reasons for taking and for not taking breaks at work. *Journal of Business and Psychology*, *38*(2), 259–282. [CrossRef] [PubMed]

Pintrich, P. R. (2000). Chapter 14—The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). Academic Press. [CrossRef]

Rheinberg, F., Vollmeyer, R., Engster, S., & Sreeeramoju, R. R. (2023). *FSS—Flow short scale (English version)*. Available online: https://www.researchgate.net/publication/373975862_FSS_-_Flow_Short_Scale_English_Version (accessed on 20 January 2025).

Rothstein, M. A., & Shoben, A. B. (2013). Does consent bias research? *The American Journal of Bioethics*, *13*(4), 27–37. [CrossRef]

Sammonds, G. M., Mansfield, N. J., & Fray, M. (2017). Improving long term driving comfort by taking breaks—How break activity affects effectiveness. *Applied Ergonomics*, *65*, 81–89. [CrossRef]

Schloemer, P., & Brenan, K. (2006). From students to learners: Developing self-regulated learning. *Journal of Education for Business*, *82*(2), 81–87. [CrossRef]

Scroggs, L. (n.d.). *De Pomodoro-techniek: Overkom uitstelgedrag en verbeter je concentratie, één pomodoro tegelijk*. Todoist. Available online: https://www.todoist.com/nl/productivity-methods/pomodoro-technique (accessed on 4 May 2025).

Sheldon, R., & Wigmore, I. (2022). *What is Pomodoro Technique time management?* Available online: https://www.techtarget.com/whatis/definition/pomodoro-technique (accessed on 4 May 2025).

Shen, M. (2023, December 31). *Flowtime technique—The best Pomodoro alternative for time management*. Flowmodor. Available online: https://flowmodor.com/blog/flowtime-technique-the-best-pomodoro-alternative-for-time-management (accessed on 27 January 2025).

Staff, C. (2025, March 25). *The Pomodoro method: Study smarter, not harder*. Coursera. Available online: https://www.coursera.org/articles/pomodoro-study-method (accessed on 1 May 2025).

Thompson, M. A., Tan, L., Soto, C., Dixit, J., & Ziat, M. (2021, October 18–22). *Tomato dice* [Conference session]. 2021 International Conference on Multimodal Interaction (pp. 628–635), Montréal, QC, Canada. [CrossRef]

Trougakos, J. P., & Hideg, I. (2009). Momentary work recovery: The role of within-day work breaks. In S. Sonnentag, P. L. Perrewé, & D. C. Ganster (Eds.), *Current perspectives on job-stress recovery* (pp. 37–84). JAI Press/Emerald Group Publishing. [CrossRef]

Willekes, N. (2014). Disrupting the flow: The detrimental effects of accelerated reader on student motivation. *Language Arts Journal of Michigan*, *29*(2), 7. [CrossRef]

Zielke, J., Anglada-Tort, M., & Berger, J. (2023). Inducing and disrupting flow during music performance. *Frontiers in Psychology*, *14*, 1187153. [CrossRef]

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, *41*(2), 64–70. [CrossRef]

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In *Handbook of metacognition in education* (pp. 299–315). Routledge/Taylor & Francis Group.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Snap & Write: Examining the Effect of Taking Photos and Notes on Memory for Lecture Content

Maribeth M. Trego [1,*], Julia S. Soares [2] and Annie S. Ditta [1]

[1] Department of Psychology, University of California Riverside, Riverside, CA 92507, USA; annie.ditta@ucr.edu
[2] Department of Psychology, Mississippi State University, Mississippi State, MS 39762, USA; js5396@msstate.edu
* Correspondence: mmcco035@ucr.edu

Abstract

Three studies investigated the effects of photo and note taking on memory for lecture content in in-person and online environments. Participants watched slideshow lecture videos and were instructed to only watch, take photos, take notes, or simultaneously take both photos and notes of the information on the slides. Memory for on-slide and only said information was tested using fill-in-the-blank questions. Experiment 1, conducted in-person, found a significant photo-taking impairment and an interaction for on-slide information such that there was a larger impairment when participants took both photos and notes compared to when they only took notes. Experiment 2 failed to replicate this interaction; there was an overall photo-taking impairment for on-slide information in an in-person learning environment. We additionally examined mind wandering as a potential mechanism driving these effects but found that it does not provide a sufficient explanation for our results. Experiment 3 used a design similar to Experiments 1 and 2 in an online environment and found a photo-taking benefit for on-slide information. Our results suggest that, in in-person classes, photo taking likely impairs learning, but, in online classes, photo taking may not be as harmful. Participants showed a note taking benefit in Experiments 2 and 3 across both class modalities.

Keywords: photo taking; note taking; education; memory

1. Introduction

In this technological age, many students take photos of lecture slides instead of or in conjunction with taking notes traditionally. Generally, research has found that note taking improves memory compared to taking no notes (Kiewra, 1989). However, most research shows that photo taking impairs memory of photographed information compared to information that was only observed (Henkel, 2014; Lurie & Westerman, 2021; Niforatos et al., 2017; Soares & Storm, 2018, 2022). That said, the effect of taking photos of lecture slides is less consistent. Some research has found no effect on memory for lectured content when participants took photos of slides compared to when they only watched the lecture, with note-taking resulting in superior memory compared to both (Wong & Lim, 2023). Other work has found a photo-taking benefit for photographed information compared to non-photographed information (Ditta et al., 2023) while others are found the reverse: impaired memory for photographed relative to non-photographed information (Wang et al., 2025). However, students often take photos and notes simultaneously, photographing some slides as they take written notes about those slides and others. Even less is known about the simultaneous effects of photo taking and note taking.

1.1. Note Taking

In general, people learn more when they take notes compared to when they do not take notes (Kiewra, 1989; Luo et al., 2018; Mueller & Oppenheimer, 2014). However, this note-taking benefit is moderated by factors such as note quality, quantity, structure, and type of question on the final test. Taking more notes with more factual statements from the learning material is often, but not always, associated with better memory performance (e.g., Kiewra, 1989; Peverly et al., 2013; Mueller & Oppenheimer, 2014; Peverly et al., 2007; Morehead et al., 2019). Taking notes that are paraphrased or summarized also tends to lead to better test performance than taking verbatim notes (Bretzing & Kulhavy, 1979; Bui et al., 2013). Additionally, Peper and Mayer (1986) found that note-taking benefits far-transfer test items significantly more than near-transfer test items. Low-ability participants particularly benefitted from not taking notes when tested for near-transfer (Peper & Meyer, 1978). As such, the effects of note taking can depend on how a student chooses to take notes, what the student is tested on, and the level of knowledge the student has in the subject at the learning phase.

1.2. Photo Taking

Taking photos of slides as a note taking strategy constitutes a particularly high-quality, high-quantity note-taking strategy, as taking photos faithfully records all of the information visually presented during a lecture. However, recording such a detailed record may not benefit memory. Given the finding that taking summarized notes is more beneficial than taking verbatim notes (e.g., Bretzing & Kulhavy, 1979), perhaps taking photos of learning material may be unhelpful or even disadvantageous to students since it does not require them to synthesize what they are learning.

Moreover, research on photo taking conducted in contexts outside of the classroom has largely demonstrated a photo-taking impairment such that memory is worse for photographed compared to non-photographed information (Henkel, 2014). The effect has been replicated using smartphone cameras (Niforatos et al., 2017; Soares & Storm, 2018, 2022). The photo-taking impairment effect has been shown using both within-subject designs comparing memory for photographed with non-photographed information and between-subjects designs in which some participants photographed all information while others photographed nothing (Soares & Storm, 2022). The impairment has also been observed on both perceptual and conceptual recognition tests, and on immediate and delayed tests (Lurie & Westerman, 2021).

There are two major proposed mechanisms argued to underlie the photo-taking impairment effect. A cognitive offloading account, first speculatively proposed by Henkel (2014), refers to the idea that photo-takers rely on their cameras to store photographed information for them. As a result, participants fail to effectively encode photographed information compared to non-photographed information. The second major proposed mechanism for the photo-taking impairment effect is that taking a photo causes global attentional disengagement to photographed information (Soares & Storm, 2018). According to this account, participants disengage generally from a photographed scene after they take a photo, and this disengagement persists even once they have put their camera away and are no longer engaging in the dual task of viewing the scene and taking a photo. Consistent with this account, the photo-taking impairment effect has been observed even when participants are given additional viewing time when they take photos to compensate for the time required to operate the camera (Henkel, 2014; Soares & Storm, 2018). An attentional disengagement account predicts that photo taking will impair memory whenever photos are taken, at least so long as attention was paid to the experience prior to taking the photo. A cognitive offloading account instead predicts that photo taking will impair memory only

when the photo-taker can rely on the camera to "remember" for them. The photo-taking impairment effect has been shown to persist even when participants know their photos will be deleted and inaccessible (Soares & Storm, 2018), a finding more consistent with attentional disengagement than cognitive offloading.

1.3. Effects of Photo Taking on Learning

Though students who take photos of educational material are unlikely to expect that they will have access to their photos on a test, prior work showing the photo-taking impairment even when photos are not expected to be reviewed suggests that similar impairments could emerge when students take photos of lecture slides. However, research on this topic is mixed.

Indeed, Ditta et al. (2023) found a photo-taking *benefit* for online recorded lecture content such that photographed information that was on the slides was remembered significantly better than non-photographed information that was on the slides, even though participants did not have the opportunity to review their photos prior to the test. This benefit was observed regardless of whether participants were instructed to take photos of certain slides or were allowed to choose which slides they wanted to photograph. This finding is consistent with previous work showing that, under some circumstances, taking photos can orient attention to a visual scene (Diehl et al., 2016). However, the photo-taking benefit observed by Ditta et al. (2023) did not always extend to information auditorily presented along with the slides by the lecturer. Prior work suggests that photo taking can have differential effects on memory for auditory and visual information (Barasch et al., 2017), so it is not particularly surprising that photo taking could differentially affect learning of lectured material that appears on-slide compared to information that is only spoken aloud by the lecturer.

Another recent study examined the effect of taking photos during a simulated in-person multimedia lecture on memory when participants did not have the ability to review their photos prior to the test. Contrary to the results of Ditta et al. (2023), Wang et al. (2025) found that participants who watched a lecture and took photos voluntarily had impaired learning compared to participants who only observed the lecture. In Wang et al.'s (2025) second experiment, participants who were assigned to take photos of every slide in a lecture scored lower on a test of their learning than participants assigned to take no photos during the lecture, an effect akin to the photo-taking impairment effect observed between subjects (Soares & Storm, 2022). However, they did not compare memory for on-slide and only-said information specifically.

Other work has also investigated the effects of taking and reviewing photos as an overall "note taking" strategy instead of comparing photographed and non-photographed information. Wong and Lim (2023) compared the effects of taking and reviewing notes on memory using three conditions: longhand note taking only, photo taking only, and a no-note/no-photo baseline. Participants who took written notes had significantly better recall for lecture content compared to participants in the photo taking and baseline conditions. Compared to baseline, participants in the photo taking condition showed no evidence of a photo-taking impairment but also no evidence of a photo-taking benefit, despite the fact that participants who took photos were able to review their photos of every slide. The authors also found that those in the note-taking condition reported less mind wandering than those in the photo taking and control conditions; thus they argued that differences in mind wandering drove the observed note taking benefit.

There is little consensus on the effect of photo taking on memory in a lecture context. That said, the extant studies compared somewhat different factors. Ditta et al. (2023) compared memory for photographed and non-photographed information that appeared

in the same lecture, similar to most photo-taking impairment studies. Others compared photo-taking as an overall strategy for recording a lecture to observing the lecture without taking photos (Wang et al., 2025) or to other strategies like traditional note taking (Wong & Lim, 2023).

It is also noteworthy that Wong and Lim (2023) and Wang et al. (2025) conducted their studies in-person, while Ditta et al. (2023) collected data entirely online. There is evidence to suggest that the attentional dynamics differ when participants are in a controlled lab environment versus online. An online participant can choose their own location, and the temptation to multitask or experience interference from outside distractors may be more prevalent online than in a controlled laboratory setting (Wang, 2022). Similar dynamics are likely to occur for remote learners. For example, self-reported attentiveness among students has been shown to vary depending on whether they attend a hybrid, synchronous, or asynchronous online course (Smith & Schreder, 2020). Ditta et al. (2023) speculated that the photo-taking benefit they observed in an online environment could be due to photo taking affecting these attentional dynamics, perhaps by reducing mind wandering or increasing attentional engagement to photographed slides. Specifically, they argued that taking a photo in a completely online context could engage more attention to the lecture compared to not taking a photo, resulting in the photo-taking benefit. As such, differences in learning format could contribute to the mixed results in the literature.

1.4. Photo and Note Taking Simultaneously

From an applied perspective, students likely use a combination of strategies, taking notes and photos of lecture slides in tandem, as they attend a lecture. Work on photo taking and its effect on memory for learned material is still developing, and, to our knowledge, no published work has yet investigated how taking both photos and notes simultaneously affects learning. There are many possible outcomes that could result from photo and note taking simultaneously in lectures. If photo and note taking both benefit memory by enhancing attention or engagement, they could have an additive relationship. In this case, taking photos and notes at the same time during a lecture would improve memory more than either method alone. Taking notes at the same time as photos could also decrease mind wandering, since a student would have to focus on managing three tasks (photo taking, note taking, and listening to the lecture) instead of only two, leaving them with little opportunity to mind-wander. However, the task-switching required to take both photos and notes while paying attention to a lecture could also become too cognitively demanding for students to juggle effectively, leading to learning impairments.

Task switching, or switching attention between two or more tasks, has been shown to induce a cost on working memory (Liefooghe et al., 2008), and students could be overwhelmed by the cognitive demands required to juggle two tasks (Draheim et al., 2016). As such, learning impairments might emerge when participants take photos and notes simultaneously that might not be observed when they engage in either task alone. Importantly, this impairment could be observed even if participants report less mind wandering when they engage in both tasks. That is, though students might mind wander less when taking photos and notes simultaneously compared to either strategy alone, the cognitive demands imposed by switching between these tasks might result in a photo-and-note-taking impairment, while engaging in either task alone might cause no such impairment or even memory benefits.

1.5. The Current Study

The present work had two main goals. First, it sought to resolve some of the inconsistencies in prior research that investigated the effects of taking photos on educational

materials. We did this by comparing memory for photographed and non-photographed information using a design similar to Ditta et al. (2023) in two supervised in-laboratory studies (Experiments 1 and 2) and one online study (Experiment 3). Second, we investigated the effects of simultaneous note and photo taking, both on memory for lecture content and on note effectiveness.

Participants watched recorded PowerPoint lectures, with slides randomly assigned to be viewed only, photographed, viewed while taking notes, or photographed and viewed while taking notes. We used a factorial design such that we could observe the effects of photo taking and note taking separately and their interaction.

Based on various findings in the literature, specifically the conflicting evidence of photo taking effects (or lack thereof) in lecture environments, the robust note-taking benefit effect, and the cognitive costs of task switching, we hypothesized that in Experiment 1, we would find a photo- and note-taking interaction effect such that information that was associated with simultaneous photo- and note-taking would be remembered worse than information that was associated with no strategy or either strategy alone. We also predicted that we would replicate the robust note-taking benefit, but, given the conflicting literature about photo taking in lecture scenarios, we were unsure whether we would observe a photo-taking benefit or impairment.

Across all experiments, as an exploratory measure, we also examined note effectiveness–a construct that encompasses different elements of the notes. Specifically, we examined the effect of simultaneous photo and note taking on note quantity: defined as overall word count; and note quality: defined as the number of correct answers to the final memory test that could be found in the notes. We investigated associations between note effectiveness and memory test performance as well as the differences in note effectiveness between the note-taking and simultaneous photo- and note-taking conditions. We hypothesized that simultaneous photo- and note-taking would result in significantly less effective notes compared to only note-taking, given that engaging in both strategies simultaneously might incur dual-task costs, which result in poorer notes.

2. Experiment 1

2.1. Participants

A total of 56 undergraduates were recruited from University of California—Riverside participant pool and were given partial course credit for their participation. A power analysis based on Ditta et al. (2023)'s observed effect size of $d = 0.41$ found that 49 participants would be required to detect a similar-sized effect with 80% power and $\alpha = 0.05$. Data from ten participants were removed for noncompliance with instructions (not taking pictures when asked, $N = 3$; taking notes when not asked, $N = 7$), for a final sample of 46. We determined that this was close enough to our stopping rule to cease data collection. Demographic information was not collected, as we did not have specific hypotheses about how the results would vary based on participant characteristics. However, we expect our sample is reflective of the general participant pool's demographics. The gender distribution in the pool is 59% female, 40% male, and 0.5% non-binary, and of the undergraduates in the pool, there are 48% freshmen, 26% sophomores, 18% juniors, and 7% seniors.

2.2. Design

A fully within-subjects 2 (photo taking: photo vs. no photo) $\times$ 2 (note-taking: notes vs. no notes) factorial design was employed. Participants watched two lecture videos under different instructions to create these four conditions. For the photo taking manipulation, participants were instructed to take pictures of a randomly selected half of the lecture slides during both lectures. For the note-taking manipulation, participants were told to take notes

throughout one lecture and to take no notes during the other lecture. Performance on a memory test, which included questions about information that appeared on the lectured slides and information that was only spoken aloud by the lecturer, was measured. Notes were also analyzed for effectiveness (i.e., quality and word count).

2.3. Materials & Measures

2.3.1. Lecture Videos

Two short lecture videos on cheesemaking and printmaking were used in this study. These topics were chosen because undergraduate psychology students would be unlikely to be familiar with the topics, and they had previously been used by Ditta et al. (2023). These videos consisted of 10 PowerPoint slides containing text and images but no animations and an auditory recording that narrated the slides. Each lecture was approximately 5 min long (range: 4:47–5:23). Each set of materials was adapted from a Wikipedia article on the subjects mentioned above by turning the text into bullet points and adding images. The narration that accompanied the slides was a section of the Wikipedia article about that topic, read in full. On all slides, there was an icon in the bottom left corner that indicated whether participants should take a photo (a camera) or not take a photo of that slide (a camera with a red slash through it).

As described above, there were two lecture types: the photo-only lecture, in which participants took photos of half of the slides and only observed the other half of the slides, and the photo + notes lecture, in which participants took longhand notes in a spiral-bound notebook throughout the lecture while simultaneously taking photos of half of the slides. These two lecture types were then broken down into four conditions for analysis: baseline (slides that were only viewed), photo only (slides that were only photographed), note only (slides for which participants only took notes), and photo + note (slides for which both notes and a photo were taken). The note-taking condition order and assignment to lecture was counterbalanced across participants.

2.3.2. Final Memory Test

Two 15-question fill-in-the-blank tests, one for each of the lectures, were adapted from Ditta et al. (2023). Sample questions include "Cheese is a food derived from milk, produced in a range of flavors, textures, and forms by coagulation of the milk protein ___" (answer: casein); and "In printmaking, each print is thought of as a(n) ___." (answer: original). These test questions were developed based on the exact text that appeared on the slides and spoken information included in the script and addressed content spread throughout the lecture. All questions for both tests can be found in the Supplementary Materials. Seven questions asked about content that appeared directly on the slides, and eight questions asked about content that was only spoken by the lecturer and did not appear on the slides.

2.4. Procedure

The entire study was conducted in the lab using a computer survey built in Qualtrics. After providing informed consent, participants were asked to put their phone on airplane mode, told that they would watch two lecture videos, and would subsequently be tested on the material. Before the photo-only lecture, they were told "*During this lecture, take a photo of each slide that has the following graphic in the bottom left corner*", (referring to the camera icon), and before the photo + note lecture, they were told "*During this lecture, please take notes on the material while also taking a photo of each slide that has the following graphic in the bottom left corner. You should write your notes in the notebook provided by the researcher. You may format the notes however you like.*". They were not informed that they would not have a period to review their notes/photos and would not have access to their photos/notes during the test. To familiarize the participants with the photo taking prompts, they were

exposed to the photo and no-photo icons before watching the lectures. For the photo taking manipulation, participants used their own smartphones to take the photos. Participants used headphones to listen to the lectures.

Participants then watched the lectures, one for which they took notes and one for which they took no notes. They could not advance to the next part of the study until the full time of the lecture had elapsed. They were watched by an experimenter to ensure that they did not interfere with the lecture (e.g., play the lectures at faster speeds, pause, rewind, or fast-forward the lectures) and that they took photos and notes as instructed. During the note-taking lecture, participants had access to a notebook and pencil in addition to their smartphone. Once the lecture was finished, the notebook and pencil were removed by the researcher.

Once the two lecture videos were complete, students completed an engaging online game, Bubble Shooter, as a distractor task for five minutes. Participants were then prompted to complete the two tests on the lectures and were not allowed to review their notes or their photos before the tests. They were given as much time as needed to complete the tests and were told to only use the knowledge in their head to complete them, so they had no access to their phones, photos, or the internet. The questions for each lecture were presented in a block-randomized order. Within the blocks, the question order was randomized. After completion of both tests, participants were debriefed, granted credit, and thanked for their participation. The entire study took under 30 min to complete.

3. Experiment 1 Results and Discussion

3.1. Memory Performance

Memory performance was scored based on the rubric found in the Supplementary Materials. Misspellings were counted as correct. Average scores per condition were calculated to represent a proportion of correct answers out of the total number of questions. A repeated-measures 2 (photo taking: photo vs. no photo) × 2 (note-taking: notes vs. no notes) factorial ANOVA found a significant main effect of photo taking, $F(1, 45) = 5.93$, $p = 0.020$, $\eta^2 = 0.02$, such that memory for photographed content ($M = 0.21$, $SD = 0.14$) was impaired relative to memory for non-photographed content ($M = 0.26$, $SD = 0.16$). We did not find a significant main effect of note-taking, $F(1, 45) = 0.77$, $p = 0.364$, $\eta^2 = 0.004$ (see Figure 1).

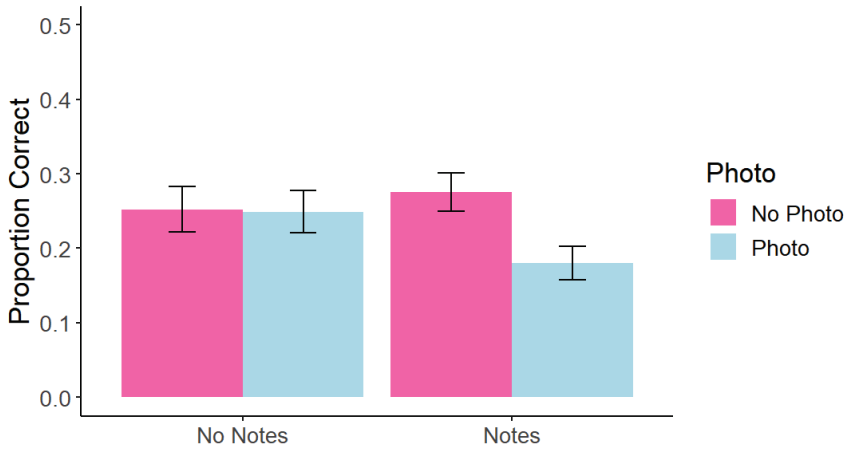


Figure 1. Experiment 1: Correct Answers at Test Across Conditions. Note. Error bars represent standard error.

As hypothesized, there was a significant interaction such that there was a larger impairment in the photo + note condition ($M = 0.18$, $SD = 0.15$) compared to the note

only condition (M = 0.28, SD = 0.17), $F(1, 45) = 4.83$, $p = 0.024$, $\eta^2 = 0.02$. To further investigate this interaction, we conducted planned comparisons. Focusing on the No Notes condition, there was no evidence of a photo-taking impairment effect $t(45) = 0.10$, $p = 0.983$. However, in the Notes condition, we found a significant photo-taking impairment effect such that performance in the No Photo condition was higher than that in the Photo condition, $t(45) = 3.97$, $p = 0.001$. These results support our hypothesis that participants in the photo + note condition would perform worse compared to the note condition, perhaps due to the high cognitive demand of multitasking between photo- and note-taking, but does not support our hypothesis that participants in the photo + note condition would perform worse compared to the photo-only condition, perhaps due to the beneficial effects of note taking.

Additionally, like Ditta et al. (2023), we compared test performance separately for information that was presented on the slides (on-slide) and information that was only said by the lecturer (only-said). A repeated-measures 2 (photo taking: photo vs. no photo) × 2 (note taking: notes vs. no notes) factorial ANOVA on the on-slide information found a significant main effect of photo taking, such that memory for photographed on-slide content (M = 0.19, SD = 0.18) was impaired relative to memory for non-photographed on-slide content (M = 0.29, SD = 0.14), $F(1, 45) = 12.37$, $p = 0.001$, $\eta^2 = 0.04$. We did not find a significant main effect of note-taking, $F(1, 45) = 0.06$, $p = 0.795$, $\eta^2 < 0.001$. There was a significant interaction, such that there was a larger impairment in the photo + note condition (M = 0.18, SD = 0.02) compared to the note only condition (M = 0.28, SD = 0.03), $F(1, 45) = 4.23$, $p = 0.046$, $\eta^2 = 0.013$ (see Figure 2). As in the analysis for the overall memory scores, we found no evidence of a photo-taking effect in the No Notes condition, $t(45) = 0.98$, $p = 0.760$, but a significant photo-taking impairment effect in the Notes condition, $t(45) = 4.52$, $p = 0.0003$. No only-said comparisons were significant; all $p > 0.076$.

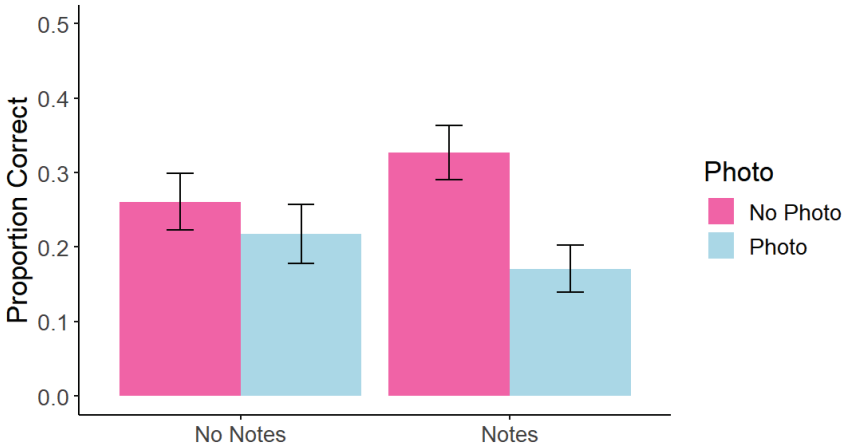


Figure 2. Experiment 1: Correct Answers at Test Across Conditions for On-Slide Information. Note. Error bars represent standard error.

3.2. Note Effectiveness

We then examined note effectiveness within the note-taking lecture. First, we found that note quantity (i.e., overall word count; M = 69.83, SD = 28.66) was not significantly correlated with test performance, $r(44) = -0.04$, $p = 0.777$.[1] Additionally, we examined note quality as the number of correct test answers found in the notes. Specifically, we compared the proportion of answers found in the notes between slides where participants only took notes versus when they were taking photos and notes simultaneously. A paired-samples t-test found no significant difference in note quality for the photo + note slides (M = 0.16, SD = 0.19) and the note only slides (M = 0.19, SD = 0.16), $t(45) = -0.98$, $p = 0.330$, $d = 0.12$.

Additionally, note quality was not significantly correlated with memory performance for either the photo + note slides, $r(44) = 0.09$, $p = 0.562$, or the note only slides ($M = 0.19$, $SD = 0.16$), $r(44) = -0.06$, $p = 0.695$.

The photo-taking impairment found in Experiment 1 when participants took photos and notes simultaneously was consistent with our predictions. It is possible that task switching between taking photos and taking notes may have induced attentional costs beyond just taking notes. Participants may have also disengaged from the main task of learning when they took photos, either during the act of photo taking, or when photographed slides were on the screen.

However, the lack of a note-taking benefit is surprising, given the often-observed benefits of note-taking (Luo et al., 2018; Mueller & Oppenheimer, 2014; Wong & Lim, 2023). The overall photo-taking impairment effect, while driven by the interaction, is in line with past photo taking research in museum contexts (Henkel, 2014), and with one classroom context study (Wang et al., 2025), but misaligned with findings from other learning contexts (Ditta et al., 2023; Wong & Lim, 2023).

4. Experiment 2

Given the somewhat contradictory findings across studies finding beneficial, null, and impairment effects of photo taking in lectures, Experiment 2 sought to replicate the effects from Experiment 1 while investigating the mechanism of the photo-taking impairment effect observed. Wong and Lim (2023) claimed that mind wandering mediates the effects of taking photos and notes, so mind wandering probes were included in Experiment 2 to test this mind wandering hypothesis. If mind wandering drives differences in learning effects of photo- and note-taking, we would expect to see the most mind wandering when participants take photos and notes simultaneously, as this condition had the worst memory performance in Experiment 1.

However, taking photos and notes simultaneously is likely the most cognitively demanding condition; if this is the case, then mind wandering rates might instead be lower in this condition compared to others. If this pattern is observed, it would suggest that something other than mind wandering, like increased cognitive demand, drives the photo-taking impairment in the note-taking conditions. Measurements of working memory capacity (WMC) were also included at the end of the study for an exploratory analysis; findings related to WMC can be found in the Supplementary Materials.

4.1. Participants

All participants were recruited using the same participant pool as Experiment 1. A total of 237 participated in the study. Data were collected for 2 quarters in an attempt to maximize the probability of finding a decisive result given that Experiment 2 included a close replication of Experiment 1. Data from 21 participants were removed for not following instructions or other technical issues ($N = 9$ for not taking pictures when asked, $N = 7$ for taking notes when told not to, $N = 4$ due to technical issues, and $N = 1$ for requesting their data be removed), for a total of $N = 216$.

4.2. Design

The same fully within-subject 2 (photo taking: photo vs. no photo) $\times$ 2 (note taking: notes vs. no notes) factorial design from Experiment 1 was employed.

4.3. Materials & Measures

All materials and procedural steps were the same as in Experiment 1, with the addition of mind wandering probes and an electronic OSPAN task (described in the Supplementary Materials; Foster et al., 2015).

Mind Wandering Probes

Mind wandering was measured using the design from Wong and Lim (2023). A total of four mind wandering probes were included (two per lecture). At two points during each lecture, one at least 30 s into the lecture and one at least 30 s before the ending of each lecture, a chime sound effect was played and the question "Are you mind-wandering?" was presented as an overlay on the slide for two seconds. Participants would then respond with yes or no on a sheet of paper separate from their notes. Participants were informed that mind wandering occurs when one's attention drifts, and they are no longer fully concentrating on the task at hand. Following the instructions used in Wong and Lim (2023), they were instructed to answer "no" if they were focused on and following the lecture content and/or activities (i.e., note and photo-taking) and "yes" in all other circumstances. No probes were presented during the transition from one slide to another, as participants were likely to be taking photos per the instructions on the new slide that appeared (if applicable).

4.4. Procedure

This study was run in Qualtrics with the same instructions and consent process as Experiment 1. However, during the instruction phase, participants were exposed to the mind wandering probe with a practice trial. The practice trial consisted of a 15 s video of a fencing lecture in the same style as the full 5 min lecture videos used in the study, accompanied by the probe overlay. Participants were also informed that their responses about their mind wandering would not affect their participation credit and that they should answer the probe honestly. They were not told how many probes would appear during each lecture.

The lectures and testing were performed following the same procedure as Experiment 1. After the memory test, participants took the electronic OSPAN task (see Supplementary Materials) and were debriefed. The entire study took under 40 min to complete.

5. Experiment 2 Results and Discussion

5.1. Memory Performance

Test data were scored as in Experiment 1. A within-subjects 2 (photo taking: photo vs. no photo) $\times$ 2 (note-taking: notes vs. no notes) factorial ANOVA found a significant main effect of photo taking, $F(1, 215) = 13.02$, $p < 0.001$, $\eta^2 = 0.005$, such that memory for photographed content ($M = 0.20$, $SD = 0.14$) was worse than for non-photographed content ($M = 0.23$, $SD = 0.14$). There was also a significant main effect of note-taking such that memory was better when participants took notes ($M = 0.23$, $SD = 0.16$) than when they did not take notes ($M = 0.20$, $SD = 0.15$), $F(1, 215) = 5.54$, $p = 0.019$, $\eta^2 = 0.009$. However, we did not replicate the interaction effect from Experiment 1, $F(1, 215) = 0.57$, $p = 0.452$, $\eta^2 < 0.001$ (see Figure 3).

As in Experiment 1, we re-ran the above analysis for on-slide and only-said information separately. We found a significant main effect of photo taking, $F(1, 215) = 6.06$, $p = 0.015$, $\eta^2 = 0.005$, such that memory for on-slide photographed content ($M = 0.20$, $SD = 0.14$) was worse than for on-slide non-photographed content ($M = 0.23$, $SD = 0.14$). There was also a significant main effect of note taking, $F(1, 215) = 18.5$, $p < 0.001$, $\eta^2 = 0.02$, such that memory for on-slide information was better when participants took notes ($M = 0.26$, $SD = 0.22$) than when they did not take notes ($M = 0.19$, $SD = 0.20$). However, we did not replicate the interaction effect from Experiment 1, $F(1, 215) = 0.36$, $p = 0.550$, $\eta^2 < 0.001$ (see Figure 4). No only-said comparisons were significant; all $p > 0.266$.

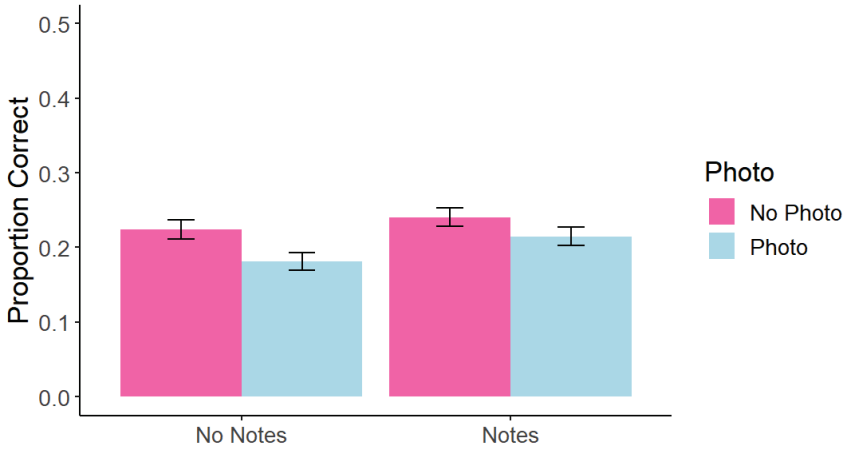


Figure 3. Experiment 2: Correct Answers at Test Across Conditions. Note. Error bars represent standard error.

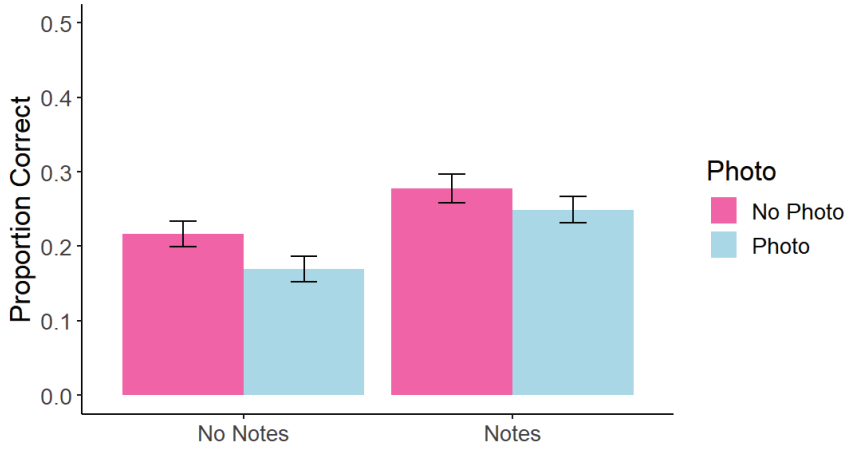


Figure 4. Experiment 2: Correct Answers at Test Across Conditions for On-Slide Information. Note. Error bars represent standard error.

5.2. Mind Wandering

Because the design of the study compared memory performance for information on different slides within (photo taking) and between (note taking) the two lectures, the same mind wandering score applied to two conditions each, based on the note-taking manipulation (i.e., the note only and photo + note conditions from the photo + note lecture vs. the photo only and baseline conditions from the photo only lecture). As such, we were not able to directly compare mind wandering as an outcome between all four conditions given that we measured it at the lecture level. Thus, a paired-samples *t*-test on mind wandering rates, ranging from 0.00 (no mind wandering at probe points), to 1.00 (mind wandering at both probe points), between the lecture with note-taking and lecture without note-taking found that participants mind wandered significantly more when they were tasked with only taking photos ($M = 0.57$, $SD = 0.64$) compared to when they were tasked with taking both photos and notes throughout the lecture ($M = 0.33$, $SD = 0.55$), $t(215) = -4.49$, $p < 0.0001$, $d = 0.40$ (see Figure 5). These findings are inconsistent with the idea that mind wandering is the sole mechanism driving the photo-taking impairment effect in lectures, as there were differing levels of mind wandering across the lectures but no evidence of a difference in the size of the memory impairment found when comparing test performance for photographed and non-photographed information across the two note-taking conditions.

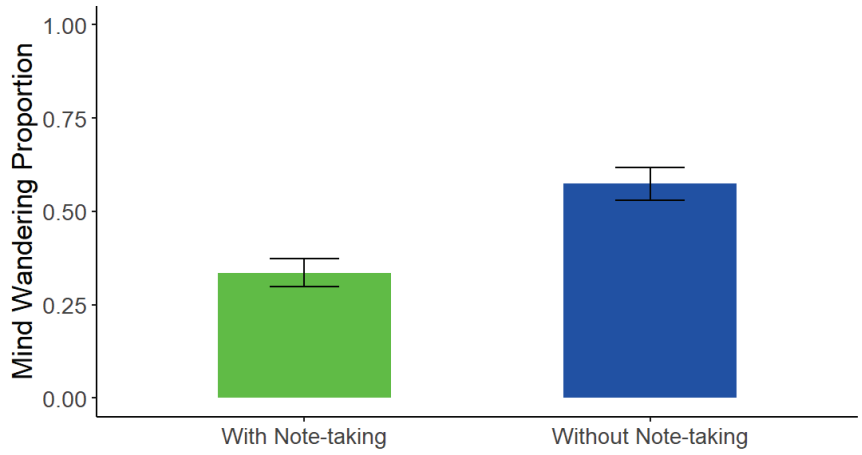


Figure 5. Experiment 2: Mind Wandering Across Lectures. Note. Error bars represent standard error.

5.3. Note Effectiveness

As in Experiment 1, note effectiveness within the note-taking lecture was examined. Consistent with Experiment 1, overall note quantity (M = 70.04, SD = 24.27) was not significantly correlated with memory performance, $r(214) = 1.43$, $p = 0.155$. However, in contrast with Experiment 1, a paired-samples t-test showed a significant difference in note quality between notes on the photo + note slides (M = 0.16, SD = 0.16) and the note slides (M = 0.20, SD = 0.17), $t(215) = -3.32$, $p = 0.001$, $d = 0.28$ (see Figure 6). Note quality was also significantly positively correlated with memory performance for both the photo + note slides, $r(215) = 0.45$, $p < 0.0001$, and the note slides, $r(215) = 0.50$, $p < 0.0001$.

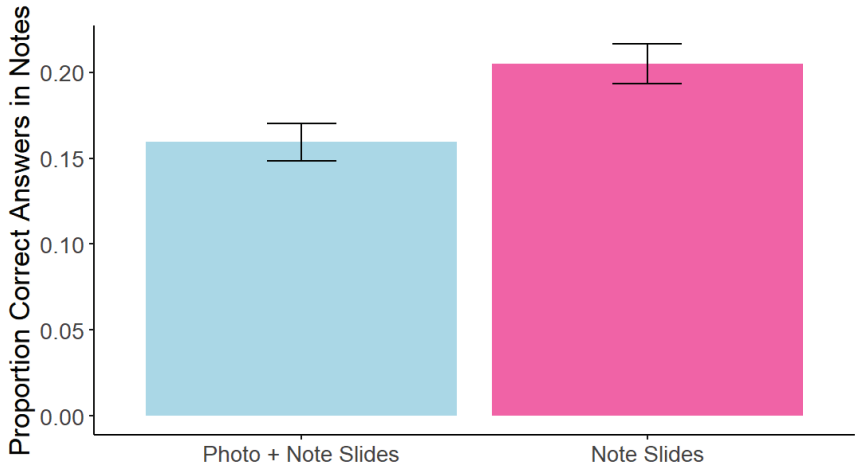


Figure 6. Experiment 2: Note Quality for Notes Taken on Photo + Note vs. Note Slides. Note. Error bars represent standard error.

Experiment 2 failed to replicate the photo- and note-taking interaction effect found in Experiment 1, but did replicate the photo-taking impairment effect, in line with Wang et al. (2025)'s findings. However, it remains unclear why the current study demonstrated a photo-taking impairment while Ditta et al. (2023), which used similar materials and procedure, found and replicated a reversal of this pattern. They found a photo taking benefit such that test performance for photographed slides was better than test performance for non-photographed slides, though the effect was more reliable for on-slide compared to only-said information. Notably, Experiments 1 and 2 were both conducted in-person, whereas Ditta et al. (2023) was conducted entirely online.

6. Experiment 3

Experiment 3 sought to determine whether an online study context would yield differential photo-taking effects; specifically, we replicated the design of the prior two experiments in an entirely online context. Attention in a remote study might be more susceptible to disruption due to mind wandering, media multitasking, and unexpected interruptions (Smith & Schreder, 2020; Wang, 2022). Thus, in a remote study, asking participants to take a photo of a slide could serve to re-engage their wandering attention on the lecture and confer a benefit not seen in in-person studies where attentional dynamics are different (i.e., participants may already be more focused on the lecture). Thus, we predicted that the photo-taking benefit effect—particularly for on-slide information—would re-emerge in an online context.

Another possibility is that Ditta et al.'s (2023) findings were anomalous. Data from their study were collected during the COVID-19 pandemic from university undergraduates who were largely taking courses online. This could have temporarily changed how these participants interacted with online lectures. Experiment 3 also investigated this possibility; the no note conditions followed Ditta et al.'s (2023) procedure with only small changes.

6.1. Participants

All participants were recruited from the same participant pool as the previous two studies. A total of 342 participants were recruited for the study. Due to the participant exclusion rate seen in Ditta et al. (2023), data was collected for 2 quarters to ensure a usable sample. Data from 150 participants were removed for not following instructions (e.g., $N = 60$ for not taking notes when asked or taking notes on both lectures; $N = 90$ for using their notes/photos during the test), for a total of 192 participants included in the analysis. This rate of participant exclusion, while high, is similar to the rate of exclusion observed by Ditta et al. (2023).

6.2. Design

The same fully within-subject 2 (photo taking: photo vs. no photo) $\times$ 2 (note-taking: notes vs. no notes) factorial design from Experiments 1 and 2 was employed. However, this experiment was run online; participants used personal computers in the location they chose to complete the study. Participants were not supervised, though compliance checks were included to ensure integrity of the data—hence the high participant exclusion rate. Outside distractors were not controlled for as participants were unsupervised; however, this mimics how students engage in remote courses as they must pay attention to the course in the face of many potential distractions.

6.3. Materials & Measures

All materials and procedural steps were the same as in Experiment 2, though the OSPAN task was removed because it could only be run using software on an in-lab computer. Additionally, a compliance check was employed to ensure participants' adherence to the study instructions.

Compliance Check

Since researchers were not present to ensure that the study instructions were followed properly, a self-report questionnaire was employed that asked about adherence to photo-, note- and test-taking instructions. Sample questions included "Please type the slide number of each slide you took a photo of in the following boxes. Slide numbers are found in the bottom right of the slide. Please separate each number with a comma. If you do not have the slide numbers in your picture, type the number of slides you took pictures of" (open-

ended answer; if participants took less than 4 or more than 5, then they were excluded); and "Did you use your notes or photos to help answer questions during the test?" (answer choices: yes, no). All questions for all tests can be found in the Supplementary Materials. Participants who reported using their notes or photos during the final memory test and those that reported taking no photos or no notes during the specified times were removed from the analysis. Participants were informed that their answers to these questions would not affect their participation credit, so they should answer honestly.

6.4. Procedure

The same procedure from Experiment 2 was employed in an online setting. However, since this was an online study and we did not have access to the participants' physical notes, the participants were asked to transcribe their notes from their notebooks into a provided box on Qualtrics at the end of this study. This allowed us to analyze their note effectiveness as in Experiments 1 and 2. They then completed the compliance check questions directly after completing the memory test but before being debriefed. This entire study took approximately 30 min to complete.

7. Experiment 3 Results and Discussion

7.1. Memory Performance

A within-subjects 2 (photo taking: photo vs. no photo) × 2 (note-taking: notes vs. no notes) factorial ANOVA found no significant main effect of photo taking such that memory for photographed content (M = 0.25, SE = 0.01) was not significantly different from non-photographed content (M = 0.23, SE = 0.01), $F(1, 191) = 3.83$, $p = 0.052$, $\eta^2 = 0.003$. There was a significant main effect of note-taking such that memory was better for lectures when participants took notes (M = 0.26, SE = 0.01) compared to when they did not take notes (M = 0.22, SE = 0.01), $F(1, 191) = 10.67$, $p = 0.001$, $\eta^2 = 0.01$. There was no significant interaction, $F(1, 191) = 0.03$, $p = 0.785$, $\eta^2 < 0.001$ (see Figure 7).

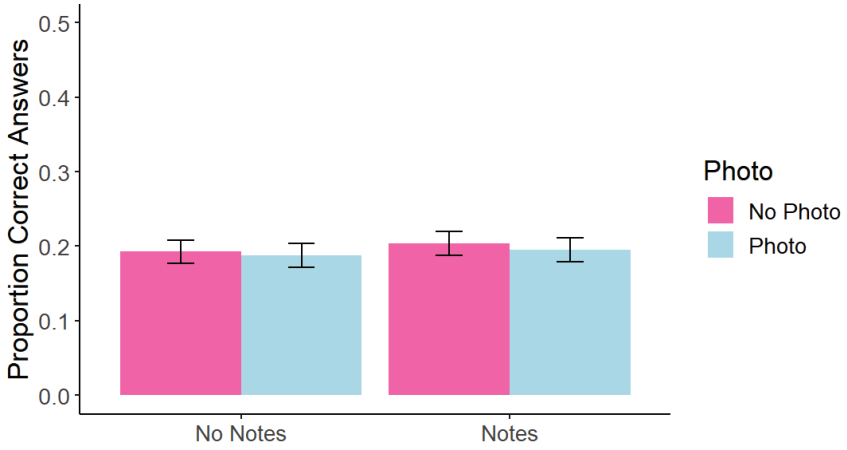


Figure 7. Experiment 3: Correct Answers at Test Across Conditions. Note. Error bars represent standard error.

However, as in the previous experiments (and in line with the analyses conducted in Ditta et al., 2023), we analyzed memory performance for the on-slide and only-said information separately. In line with Ditta et al. (2023), a within-subjects 2 (photo taking: photo vs. no photo) × 2 (note-taking: notes vs. no notes) factorial ANOVA found a significant main effect of photo taking, $F(1, 191) = 6.90$, $p = 0.009$, $\eta^2 = 0.008$, such that memory for on-slide photographed content (M = 0.31, SE = 0.002) was better than for on-slide non-photographed content (M = 0.26, SE = 0.002). There was also a significant

main effect of note taking, $F(1, 191) = 17.06$, $p < 0.001$, $\eta^2 = 0.16$, such that memory was better when participants took notes on on-slide information ($M = 0.32$, $SE = 0.02$) than when they did not take notes ($M = 0.24$, $SE = 0.02$). However, we again did not replicate the interaction effect from Experiment 1, $F(1, 191) = 0.57$, $p = 0.451$, $\eta^2 < 0.001$ (see Figure 8). No only-said comparisons were significant; all $p > 0.516$.

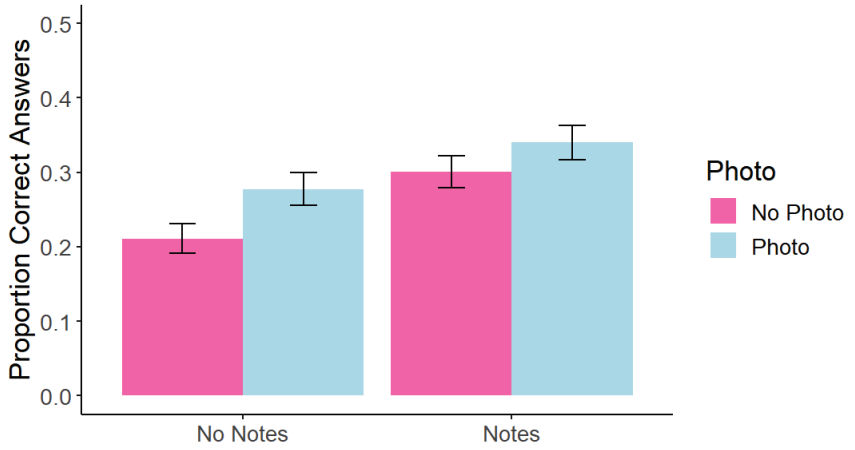


Figure 8. Experiment 3: Correct Answers at Test Across Conditions for On-Slide Information. Note. Error bars represent standard error.

7.2. Mind Wandering

A paired-samples *t*-test on mind wandering rates between the lectures did not find a significant difference between when they were only taking photos ($M = 0.54$, $SD = 0.63$) compared to when they were taking photos and notes simultaneously ($M = 0.49$, $SD = 0.66$), $t(217) = -4.49$, $p = 0.433$ (see Figure 9). This is in contrast to Experiment 2, where mind wandering was different across lectures in an in-person context–here we see high levels of mind wandering in both lectures, which is consistent with our proposal that there is less attentional engagement in online environments.

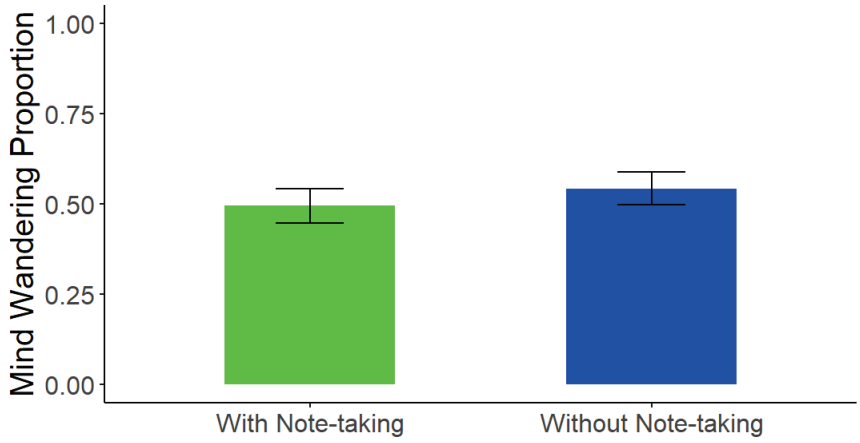


Figure 9. Experiment 3: Mind Wandering Across Lectures. Note. Error bars represent standard error.

7.3. Note Effectiveness

Inconsistent with Experiments 1 and 2, note quantity ($M = 76.83$, $SD = 55.6$) was significantly positively correlated with memory performance, $r(190) = 0.18$, $p = 0.011$. Inconsistent with Experiment 2, a paired-samples *t*-test showed no significant difference in note quality between the photo + note slides ($M = 0.19$, $SD = 0.19$) and the note slides ($M = 0.20$, $SD = 0.19$), $t(191) = -0.89$, $p = 0.373$, $d = -0.06$ (see Figure 10). Note quality was

significantly positively correlated with memory performance for both the photo + note slides, $r(190) = 0.27$, $p = 0.0002$, and the note slides, $r(190) = 0.50$, $p < 0.0001$. This means that there was no difference in note quality when taking photos and notes versus only notes and that having high quality notes was associated with better memory regardless of condition. Since note quantity is positively related to performance in an online setting, this could mean that participants who wrote more put more effort and attention towards the experiment. The overall inconsistency in note effectiveness findings across experiments may be due to differing levels of attention or effort in an online study compared to an in-person study.

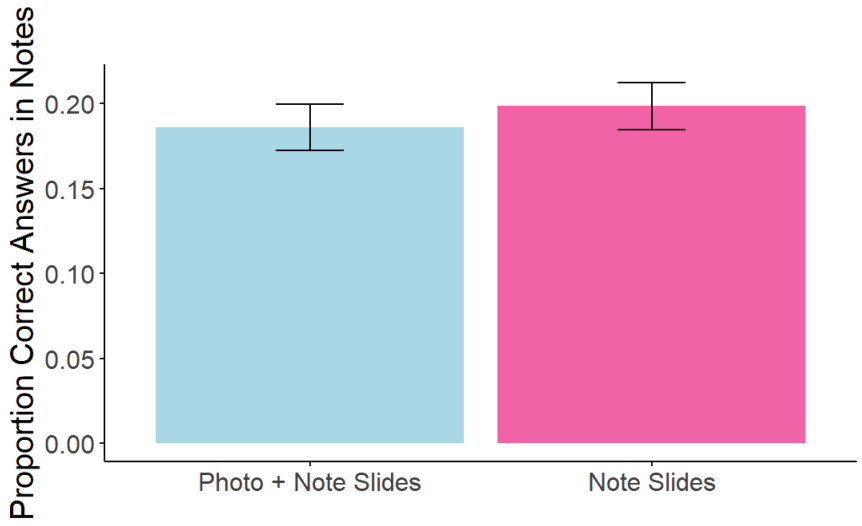


Figure 10. Experiment 3: Note Quality Between Notes Taken on Photo + Note and Note Slides. Note. Error bars represent standard error.

8. General Discussion

In three experiments, we examined how taking photos and notes simultaneously affects memory for lecture content (see Table 1); to our knowledge, this is the first investigation of the effects of this practice on learning. In Experiment 1, we observed a learning impairment associated with simultaneous photo-and note-taking, suggesting that combining these strategies negatively affects memory more than using either strategy alone. However, we failed to replicate this interaction effect in Experiments 2 and 3. In Experiment 2, we found a photo-taking impairment and a note-taking benefit in an in-person learning environment. The effects in Experiments 1 and 2 held when examining memory for on-slide content specifically, though there were no significant effects for only-said information. In Experiment 3, which was conducted in an online context, we found a note-taking benefit, but more importantly, we also found evidence of a photo-taking benefit for on-slide information but not only-said information, replicating the findings of Ditta et al. (2023).

Table 1. Memory Effects Across Experiments.

Exp	Environment	Photo Overall	Photo On-Slide	Note	Interaction
1	In-person	Detrimental	Detrimental	*n.s.*	Yes
2	In-person	Detrimental	Detrimental	Beneficial	*n.s.*
3	Online	*n.s.*	Beneficial	Beneficial	*n.s.*

Note. 'Photo' indicates whether a statistically significant effect was found and the direction of the effect for photo taking on memory. 'Note' indicates whether a statistically significant effect was found and the direction of the effect for note-taking on memory. Interaction indicates whether a statistically significant effect was found. Environment refers to whether the experiment was conducted in person or online.

The results of Experiment 1 and 2 align with the majority of past photo taking research that identified a photo-taking impairment effect (e.g., Henkel, 2014; Soares & Storm, 2018) and one study on photo taking in a learning context (Wang et al., 2025). These findings do not align, however, with two previous studies on photo taking in learning contexts (Ditta et al., 2023; Wong & Lim, 2023). However, the results of Experiment 3 align with previous work identifying a photo-taking benefit for on-slide information in an online context (Ditta et al., 2023).

The photo-taking impairment observed in Experiments 1 and 2 could be driven by some factor that is likely to differ across in-person and online settings, like attentional engagement. Participants' attentional engagement could be higher when watching lectures in a lab setting where they are being monitored compared to participating in an online study where they are not. Indeed, our participant exclusion rates and mind wandering findings may suggest that in-person participants were more engaged than online participants. Our findings from Experiments 1 and 2 suggest that in an in-person learning context, memory for photographed information is impaired relative to non-photographed information. These findings, coupled with those of Wang et al. (2025), suggest that, overall, taking photos in in-person settings may be more harmful than helpful for learning. The photo-taking benefit for on-slide information presented online from Experiment 3 and Ditta et al. (2023) suggest that photo taking may be less detrimental for learning in an online environment. Together, these findings preliminarily suggest that the effect of photo taking on memory for lecture content may depend on the environment, with impairments occurring in in-person (i.e., lower distraction learning scenarios) and benefits occurring online (i.e., distractor heavy learning scenarios), potentially due to differences in attentional engagement across learning modalities (Smith & Schreder, 2020; Wang, 2022). However, this possibility is speculative because it is based on cross-experiment comparisons. Future work should more directly address the possibility that the effects of photo taking on learning from lecture slides depend on one or many of the factors that differ between an in-person and remote learning environment.

Experiments 2 and 3 align with the majority of past note-taking research that shows taking notes benefits memory (Peverly et al., 2013; Mueller & Oppenheimer, 2014; Jansen et al., 2017). Previous research has suggested that lowered rates of mind wandering when participants take notes compared to when they take photos or only watch a lecture may cause these memory benefits (Wong & Lim, 2023). In Experiment 2, we replicated Wong and Lim's (2023) finding that note-taking participants report less mind wandering compared to when they did not take notes. This finding highlights the importance of note taking as a way to direct attention to a learning task (Jansen et al., 2017). However, in Experiment 3, there was no difference in mind wandering between these conditions. This may be because participants in Experiment 3 engaged less attention overall than their in-person counterparts, though this is speculative (a cross-experiment comparison of overall mind wandering rates between in-person and online environments was non-significant).

One of the goals of this work was to investigate mind-wandering as a mechanism underlying photo and note taking effects. Wong and Lim (2023) argued that mind wandering mediated the beneficial effect of note taking on memory, with the note-taking condition having significantly less mind wandering and thus higher memory performance than the photo taking and baseline conditions. Thus, we hypothesized that there would be lower mind wandering in the photo + note lecture compared to the photo-only lecture because participants would be switching between more tasks. Indeed, this is what we found in Experiment 2. However, both conditions showed a similar photo-taking impairment. Thus, increased mind wandering cannot be the sole reason behind the photo-taking impairment effect. However, these mind wandering findings do not rule out the possibility that at-

tention is involved in driving some of these effects. If attention is involved, it was not apparent in our subjective measures of mind wandering, but it could be apparent using other measures.

Our final, exploratory goal of this project was to examine how note effectiveness is affected by photo and note taking, and how memory performance and note effectiveness are related. Interestingly, the pattern of results was different across experiments. Experiment 1 had no significant note effectiveness findings; this could be due to a power issue, as the power needed was based on photo taking, not note effectiveness, effect sizes. For Experiment 2, which had more participants, we found that participants' note quality was significantly better when they focused only on taking notes compared to when they engaged in simultaneous photo and note taking. This finding aligned with our prediction that less task switching would be associated with higher note quality. However, in Experiment 3, which was online, there was no significant difference in note quality across the note taking and photo + note conditions. This finding could be due to less attention being devoted to the assigned tasks (photo taking, note taking, and watching the lecture) in the online environment, as shown by the numerically, but not significantly, higher overall rates of mind wandering in Experiment 2 compared to Experiment 3 (though again, cross-experiment comparisons should be interpreted with caution). The general inconsistency of note quantity and its relationship to test performance across the three experiments is not surprising, as the quality and content of the notes are more reliably linked to memory performance than quantity (Bretzing & Kulhavy, 1979; Bui et al., 2013). The relationship of note quality and test performance was found in Experiment 2 and 3 but not Experiment 1; this could be due to this study being underpowered for note quality analysis.

In Experiments 2 and 3, we replicated prior note taking literature (Kiewra, 1989; Peverly et al., 2013), such that higher note quality was significantly and positively correlated with better memory. However, note quantity was only significantly related to memory in an online context; this positive correlation may be due to the attentional dynamics at play in an online environment.

9. Limitations and Future Work

One of the major limitations of this series of experiments was the design of the conditions. Because we manipulated photo and note taking to be entirely within participants to increase power, it was not possible to make some comparisons that may have been useful. Specifically, we could not make direct comparisons of mind wandering across our four conditions because we measured mind wandering at the lecture level, not the condition level. Future work could remedy this by assigning four separate lectures—each with one strategy—rather than the two we used. Additionally, future work should strive to measure attention directly in an online versus in-person design so direct claims about attention and context can be made.

The current study used similar materials and procedures across experiments and compared to Ditta et al. (2023), finding both evidence of the photo-taking benefits observed by Ditta and colleagues as well as evidence for a photo-taking impairment effect. Together, these findings suggest that differences in methods, materials, or participant pools are not the factors driving discrepancies across the literature regarding the effects of photo taking on learning of online lecture content. Since we did not directly compare photo and note taking in in-person and online learning contexts within a single study, this speculation relies on cross-experiment comparisons. However, the pattern of fundings suggests that one of the many factors that differ between in-lab and remote learning could play a role in determining the direction of photo-taking effects on learning. To address this possibility, future work should manipulate class context within a single study, as well as systematically

investigate how factors that differ between these settings like accountability, distraction, and media multitasking interact with the effects of photo taking.

It is also important to note that this paradigm does not completely replicate learning in realistic classroom situations. For example, our paradigm employs an immediate memory test, rather than a delayed test that would be used in most courses. Past photo taking research has shown a persistent photo taking impairment effect from immediate to delayed testing for both factual and conceptual information (Lurie & Westerman, 2021), but that study was conducted using art objects and not educational materials, so perhaps there are differences that may emerge when using a delayed test for educational materials. An additional consideration is that students may use their photos in conjunction with study strategies (summarizing, rewriting notes) in order to prepare for a delayed test, which could negate the photo-taking impairment. According to Wong and Lim (2023) a photo review period does not significantly improve memory on an immediate test for photo taking compared to reviewing notes, but a review period with a delayed test has not been examined in a learning context.

To address these limitations, future in-lab research should seek to more closely replicate real-life synchronous in-person classroom and online learning environments, where distractions, fatigue, and motivation could be different from a lab/recorded video setting; these moderating variables should be explicitly examined. Finally, future work should consider examining the ways in which individual differences between students (e.g., English language learner status, age, academic discipline, education level, etc.) interact with photo and note taking.

10. Conclusions

Overall, this was the first study to examine simultaneous photo- and note- taking behaviors in a lecture context and their subsequent effect on memory. Though more work remains to be carried out to fully characterize the simultaneous effects of photo and note taking in the classroom and to understand the underlying mechanisms behind these effects, our results suggest that students should avoid taking photographs of slides in an in-person context; however, if in an online context, photo taking could benefit memory for on-slide information.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/bs15050561/s1, Test Questions and Answers; Working Memory Snap and Write.

Author Contributions: Conceptualization, J.S.S., M.M.T. and A.S.D.; methodology, J.S.S., M.M.T. and A.S.D.; formal analysis, M.M.T.; resources, A.S.D. and J.S.S.; data curation, M.M.T.; writing—original draft preparation, M.M.T.; writing—review and editing, A.S.D., M.M.T. and J.S.S.; visualization, M.M.T.; project administration, M.M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of University of California Riverside (HS #19-281, approval obtained 12 June 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: Data for these studies can be found here: https://github.com/tregomary/Snap-and-Write (accessed on 14 February 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

[1] Because both the photo + note and note conditions were created by comparing memory performance for information on different slides within the same lecture, the same set of notes applied to both conditions. As such, we were not able to directly compare word count (i.e., note quantity) as an outcome between the photo + note and note conditions.

References

Barasch, A., Diehl, K., Silverman, J., & Zauberman, G. (2017). Photographic memory: The effects of volitional photo taking on memory for visual and auditory aspects of an experience. *Psychological Science*, *28*(8), 1056–1066. [CrossRef] [PubMed]

Bretzing, B. H., & Kulhavy, R. W. (1979). Notetaking and depth of processing. *Contemporary Educational Psychology*, *4*(2), 145–153. [CrossRef]

Bui, D. C., Myerson, J., & Hale, S. (2013). Note-taking with computers: Exploring alternative strategies for improved recall. *Journal of Educational Psychology*, *105*(2), 299–309. [CrossRef]

Diehl, K., Zauberman, G., & Barasch, A. (2016). How taking photos increases enjoyment of experiences. *Journal of Personality and Social Psychology*, *111*(2), 119–140. [CrossRef] [PubMed]

Ditta, A. S., Soares, J. S., & Storm, B. C. (2023). What happens to memory for lecture content when students take photos of the lecture slides? *Journal of Applied Research in Memory and Cognition*, *12*(3), 421–430. [CrossRef]

Draheim, C., Hicks, K. L., & Engle, R. W. (2016). Combining reaction time and accuracy: The relationship between working memory capacity and task switching as a case example. *Perspectives on Psychological Science*, *11*(1), 133–155. [CrossRef]

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*(2), 226–236. [CrossRef]

Henkel, L. A. (2014). Point-and-shoot memories: The influence of taking photos on memory for a museum tour. *Psychological Science*, *25*(2), 396–402. [CrossRef]

Jansen, R. S., Lakens, D., & IJsselsteijn, W. A. (2017). An integrative review of the cognitive costs and benefits of note-taking. *Educational Research Review*, *22*, 223–233. [CrossRef]

Kiewra, K. A. (1989). A review of note-taking: The encoding-storage paradigm and beyond. *Educational Psychology Review*, *1*, 147–172. [CrossRef]

Liefooghe, B., Barrouillet, P., Vandierendonck, A., & Camos, V. (2008). Working memory costs of task switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(3), 478–494. [CrossRef] [PubMed]

Luo, L., Kiewra, K. A., Flanigan, A. E., & Peteranetz, M. S. (2018). Laptop versus longhand note taking: Effects on lecture notes and achievement. *Instructional Science*, *46*, 947–971. [CrossRef]

Lurie, R., & Westerman, D. L. (2021). Photo-taking impairs memory on perceptual and conceptual memory tests. *Journal of Applied Research in Memory and Cognition*, *10*(2), 289–297. [CrossRef]

Morehead, K., Dunlosky, J., & Rawson, K. A. (2019). How much mightier is the pen than the keyboard for note-taking? A replication and extension of Mueller and Oppenheimer (2014). *Educational Psychology Review*, *31*, 753–780. [CrossRef]

Mueller, P. A., & Oppenheimer, D. M. (2014). The Pen Is Mightier Than the Keyboard: Advantages of Longhand Over Laptop Note Taking. *Psychological Science*, *25*(6), 1159–1168. [CrossRef]

Niforatos, E., Cinel, C., Mack, C. C., Langheinrich, M., & Ward, G. (2017). Can less be more?: Contrasting limited, unlimited, and automatic picture capture for augmenting memory recall. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(2), 1–22. [CrossRef]

Peper, R. J., & Mayer, R. E. (1978). Note taking as a generative activity. *Journal of Educational Psychology*, *70*(4), 514–522. [CrossRef]

Peper, R. J., & Mayer, R. E. (1986). Generative effects of note-taking during science lectures. *Journal of Educational Psychology*, *78*(1), 34–38. [CrossRef]

Peverly, S. T., Ramaswamy, V., Brown, C., Sumowski, J., Alidoost, M., & Garner, J. (2007). What predicts skill in lecture note taking? *Journal of Educational Psychology*, *99*(1), 167–180. [CrossRef]

Peverly, S. T., Vekaria, P. C., Reddington, L. A., Sumowski, J. F., Johnson, K. R., & Ramsay, C. M. (2013). The relationship of handwriting speed, working memory, language comprehension and outlines to lecture note-taking and test-taking among college students. *Applied Cognitive Psychology*, *27*(1), 115–126. [CrossRef]

Smith, J., & Schreder, K. (2020). Are they paying attention, or are they shoe-shopping? Evidence from online learning. *International Journal of Multidisciplinary Perspectives in Higher Education*, *5*(1), 200–209. [CrossRef]

Soares, J. S., & Storm, B. C. (2018). Forget in a flash: A further investigation of the photo-taking-impairment effect. *Journal of Applied Research in Memory and Cognition*, *7*(1), 154–160. [CrossRef]

Soares, J. S., & Storm, B. C. (2022). Does taking multiple photos lead to a photo-taking-impairment effect? *Psychonomic Bulletin & Review*, *29*(6), 2211–2218.

Wang, C. (2022). Comprehensively summarizing what distracts students from online learning: A literature review. *Human Behavior and Emerging Technologies*, *2022*(1), 1483531. [CrossRef]

Wang, C., Li, X., & Jin, H. (2025). Photograph taking in the classroom impairs memory of learned material. *Social Behavior and Personality: An International Journal*, *53*(1), 1–8. [CrossRef]

Wong, S. S. H., & Lim, S. W. H. (2023). Take notes, not photos: Mind-wandering mediates the impact of note-taking strategies on video-recorded lecture learning performance. *Journal of Experimental Psychology: Applied*, *29*(1), 124–135. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

behavioral sciences

MDPI

Article

Examining the Potential Benefits of Affirming Values on Memory for Educational Information

Karen Arcos *, Rebecca Covarrubias and Benjamin C. Storm

Department of Psychology, University of California, Santa Cruz, Social Sciences 2, Room 277, Santa Cruz, CA 95064, USA; rgcovarr@ucsc.edu (R.C.); storm@ucsc.edu (B.C.S.)
* Correspondence: karcos1@uci.edu

Abstract

First-generation students can experience a cultural mismatch between their values and those that colleges and universities tend to prioritize. This mismatch can increase cognitive load, leaving fewer resources available for learning. Effective and long-lasting learning requires actively processing new information and connecting it to existing knowledge—an effort that demands significant cognitive resources. Value affirmation exercises, where students select and reflect upon values that are important to them, have shown promise in reducing cultural mismatch and improving performance on cognitive tasks. However, the impact of these exercises on the learning and recall of new information is less clear. The current study investigated whether a value affirmation exercise, completed before reading an educational passage, would improve memory recall for that passage in a sample of 400 first-generation and continuing-generation young adults, as compared to not affirming. Our results failed to provide evidence that value affirmation exercises impacted recall performance, regardless of whether participants affirmed independent values, interdependent values, or both. Given the importance and implications of this outcome for student learning, we discuss possible explanations for these null findings and suggest future directions in affirmation research.

Keywords: learning; memory; value affirmation interventions; first-generation college students; culture; cultural mismatch; independence; interdependence

1. Introduction

U.S. institutions of higher education reward cultural values and practices rooted in independence and individualism, such as pursuing individual success and competition (Goudeau et al., 2025; Phillips et al., 2020; Stephens et al., 2012a). Students socialized in contexts that affirm similar values as compared to the university, like middle-to-upper-class contexts, can experience the university as a familiar cultural setting. This cultural match may enable these students to dedicate more of their cognitive resources toward learning new and potentially challenging information. However, students socialized in contexts affirming interdependent values (e.g., connecting to others, valuing cooperation over competition), like working-class contexts, can experience the university's independent culture as a mismatch (Covarrubias et al., 2019; Phillips et al., 2020; Stephens et al., 2012a; Valle & Covarrubias, 2024; Vasquez-Salgado et al., 2015).

According to cultural mismatch theory, the university norms of independence advantage students from middle-to-upper-class and continuing-generation college backgrounds, while marginalizing students from first-generation college backgrounds, a difference that can lead

to disparities in educational outcomes (Phillips et al., 2020; Stephens et al., 2012a). When exposed to the same university message of independence, for example, first-generation students have been shown to cognitively underperform compared to continuing-generation students (Stephens et al., 2012a). Several factors may drive this effect. When first-generation students experience cultural mismatch, they tend to experience increased stress (i.e., cortisol levels), negative emotions, and heightened perceptions of task difficulty compared to what they would have experienced otherwise (Stephens et al., 2012b). Aligning university norms and values with interdependence, however, can ameliorate such effects (Hecht et al., 2021; Tibbetts et al., 2016).

Indeed, cultural motivations for learning—like independent or interdependent motivations—shape what information students pay attention to and/or what they find to be informative. Framing cues from the learning environment to match or align with how learners perceive their environment can affect how students encode, store, and recall to-be-learned information (Wang, 2009). These cultural elements also impact how people conceptualize aspects like space and time, which can in turn affect how events in episodic memory are segmented and ultimately recalled (see Gutchess & Rajaram, 2023; Wang, 2021 for reviews on cultural influences on human memory and cognition).

1.1. Signaling a Cultural Match: Value Affirmation Exercises

In contexts where dominant cultural messages persistently mismatch students' values and motivations, light-touch yet meaningful strategies for signaling a cultural match can support students in their learning and educational outcomes. One such intervention is the use of value affirmation exercises. Value affirmation exercises ask people to choose from a select list of values that are important to them and to reflect on why those values matter to them (Steele, 1988). Across a series of studies, such exercises have been shown to boost performance and feelings of belonging of marginalized students compared to a control condition in which participants select and reflect on values least important to them but important to someone else (Borman et al., 2021; Cohen et al., 2006, 2009; Harackiewicz et al., 2016; Hecht et al., 2021; for a meta-analysis, see Wu et al., 2021). Value affirmation interventions are believed to be effective in part because they enable students to regain or strengthen a sense of self-worth in a context that consistently threatens or stigmatizes aspects of their identities.

A subset of the literature on value affirmation interventions has focused on understanding how independent and/or interdependent values function to affirm students' identities (Covarrubias et al., 2016; Hoshino-Browne et al., 2005; Tibbetts et al., 2016). For example, one study tested the long-term benefits of a value affirmation intervention in a large introductory biology course (Tibbetts et al., 2016). As part of a class assignment, participants in the intervention condition chose and reflected on independent and interdependent values that were important to them. First-generation students who affirmed their values earned significantly higher grades both that semester and over the following three years compared to first-generation students who did not. Follow-up analyses showed that the extent to which participants wrote about independent values primarily explained this effect. That is, when first-generation students affirmed values that matched the classroom and university culture of independence, their grades improved in subsequent semesters. In this way, value affirmation strategies may function as a pathway for creating a match with university culture.

In a second study, the same authors explored the impacts of affirming independent values on students' performance on GRE-style math problems (see also Miyake et al., 2010; Tibbetts et al., 2016; Yeager & Walton, 2011). First-generation students who affirmed and reflected on their independent values significantly outperformed first-generation students who affirmed and reflected on their interdependent values. The authors argued

that, in affirming values related to independence, students might be able to perceive the learning environment as less socially threatening, thereby allowing them to attend more fully to the challenging cognitive task at hand (Tibbetts et al., 2016).

Importantly, affirming interdependent values has also been shown to benefit marginalized students. For example, when Latinx university students engaged in an interdependent value affirmation exercise (i.e., affirming values important to them and their families), they outperformed Latinx students who engaged in a more independent value affirmation exercise (i.e., affirming values that were individually important to them) on a word puzzle anagram task (Covarrubias et al., 2016). Similarly, other studies have found that women affirming interdependence performed better on math tests, possibly because thinking of their connections with others bolstered their sense of identity and, thus, reduced the cognitive impact that experiencing threat can cause (Shnabel et al., 2013). So, affirming independent or interdependent values within a university setting can both signal a cultural match and, separately, facilitate positive outcomes.

As would be expected, given their distinct capacity to produce positive benefits, affirming both independence and interdependence in a combined value affirmation exercise has also been shown to positively impact academic performance. For example, Tibbetts et al. (2018) found that first-generation students at two-year colleges earned higher course grades when they affirmed both interdependent and independent values compared to a control condition. Moreover, first-generation students from racially minoritized backgrounds (e.g., Latinx, Black) earned higher grades when affirming independence compared to the control condition, whereas the opposite was true for first-generation students from racial majority backgrounds (e.g., White). No differences were found for interdependent-related affirmations. These findings reveal how affirming independence as well as both independence and interdependence can impact learning.

Because the effects and experiences related to cultural mismatch may be less pronounced for participants attending two-year colleges compared to those attending four-year universities (Stephens et al., 2012a; Tibbetts et al., 2016), Hecht et al. (2021) sought to examine the impact of a combined intervention for first- and continuing-generation students attending a four-year university. Again, they found that, compared to a control condition, affirming both interdependence and independence yielded higher scores on a math test and on a cultural match measure among both first- and continuing-generation students. The value affirmation task appeared to reduce participants' cognitive load and increase confidence amidst stress.

1.2. Current Study Aims

In summary, considerable evidence now supports the positive impacts of value affirmation interventions in educational settings. Exactly how such interventions affect learning, however, remains to be fully understood. Moreover, researchers have observed improved performance on specific cognitive tasks (e.g., solving math problems and word puzzles) following an affirmation exercise, as well as in long-term outcomes such as student grade point average. However, little direct research exists on the impact of affirmation exercises on tasks designed to simulate the kind of learning commonly experienced in the classroom. As such, one of the primary goals of the current study was to examine the impact of a value affirmation exercise on students' ability to read and remember an educational passage.

Research has shown that the ways in which learners engage with to-be-learned information can significantly impact their memory and the extent to which that information becomes integrated with information and knowledge entrenched in long-term memory (e.g., Agarwal & Roediger, 2018; Pan & Bjork, 2024; Rajaram & Barber, 2024). Moreover, learners often employ suboptimal strategies, with students spending substantial time and

effort engaging in activities that may feel like they are producing learning even though they are not (Dunlosky et al., 2013). For example, learners often avoid challenging experiences (e.g., Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007), despite the potential of such experiences to enhance long-term learning and transfer (Bjork & Bjork, 2011).

In the context of reading and learning an educational passage, students have significant control in determining the cognitive processes in which they engage. For example, students may or may not retrieve what they had previously read across a given passage to integrate information, thereby synthesizing the content and producing new insights and ideas that further support memory. Likewise, students may or may not interrogate or question what they know, what they are reading, or connect the content to their own interests and existing knowledge. Students who engage in these kinds of processes are much more likely to remember a passage's contents than those who do not.

Whether students are able to engage in and benefit from these kinds of "desirable difficulties" during self-directed learning is likely to depend on the availability of cognitive resources such as working memory and executive control (e.g., Cortés Pascual et al., 2019). Students who experience cultural mismatch, and other threats related to their identity, may have fewer cognitive resources to engage in the kinds of processes and strategies that benefit learning. If students must allocate cognitive resources towards managing the psychological feelings of stress and negative emotions that cultural mismatch induces, then they may apply fewer cognitive resources towards effectively implementing valuable learning strategies. Similar constraining effects have been demonstrated in a variety of ways in the context of stereotype threat (e.g., Schmader et al., 2008; Szmalec et al., 2011).

Value affirmation exercises may reduce stress and negative emotions, thus mitigating the negative impacts of cultural mismatch. This mitigation allows the cognitive resources that would have been expended towards managing such experiences to be reallocated towards learning. Moreover, value affirmation exercises could encourage learners to engage more deeply with to-be-learned material. In the present study, participants at four-year universities (consisting of both first-generation and continuing-generation students and recent graduates) were randomly assigned to one of four conditions: an independent condition in which they affirmed independent values; an interdependent condition in which they affirmed interdependent values; a combined condition in which they affirmed both independent and interdependent values; or a control condition in which they did not affirm any values. Participants were then asked to read an educational passage and then attempt to recall as much information from the passage as possible.

Unlike past studies' control conditions, which instructed participants to affirm the values least important to them (Borman et al., 2021; Cohen et al., 2006, 2009; Harackiewicz et al., 2016; Hecht et al., 2021), we utilized a no-value-affirmation control condition. This was because some past studies instructing participants to reflect on values "important to someone else" have inadvertently operated as interdependent affirmations, especially when participants reflect on values important to family (Covarrubias et al., 2016). To avoid the potential for this, we opted to use a no-value-affirmation control condition, in which participants were asked to find anagrams amidst scrambled letters, a task that past value affirmation studies have employed (Covarrubias et al., 2016; Stephens et al., 2012a), albeit as a dependent variable instead of as a control condition. The anagram task was designed to engage participants and require their attention in a way similar to the affirmation conditions without creating stress, threat, or affecting their sense of belonging.

Through this design, we aimed to make at least two contributions. First, to our knowledge, no studies published to date have measured the immediate effects of a value affirmation exercise on participants' recall of newly learned information. Second, by including a combined condition, we sought to better tease apart the singular and

additive effects of independent and interdependent affirmations for students attending or who have attended a four-year university. Based on the research reviewed above, we expected first-generation participants to benefit from conditions that involved affirming independent values (Miyake et al., 2010; Tibbetts et al., 2016; Yeager & Walton, 2011), as well as conditions that involved affirming interdependent values (Covarrubias et al., 2016; Hoshino-Browne et al., 2005), perhaps accruing the greatest benefits when affirming both (Tibbetts et al., 2018). Owing to the lack of pre-existing cultural mismatch, we expected the affirmation intervention to affect continuing-generation students relatively less. More generally, the study provides an important step towards better understanding whether and how different kinds of value affirmation interventions have the potential to impact learning.

2. Methods

2.1. Participants

Prior to recruitment, we conducted a power analysis to gain the thresholds for 80% power. Each condition needed 100 participants to achieve a Cohen's *d* of 0.40 between any two conditions. Thus, we aimed to recruit 400 participants and stopped data collection as soon as that target was met. This sample size was chosen in an effort to best balance our resource constraints and desire to have enough power to detect a small to medium effect size that would be both practically and theoretically informative.

A total of 400 participants completed the study (*Mean* Age = 23.11; *SD* = 2.63; range = 18 to 34 years). Participants were recruited using Prolific software (Prolific, London, UK), 1 November–1 December 2022. Participants were eligible if they were 18–27 years of age (four participants were included despite being older than the maximum age), resided in the U.S., and were either current undergraduate students or alumni who had recently earned their undergraduate degrees from four-year institutions. Students from all majors were eligible. To recruit participants, we prescreened for age, English fluency, and an approval rate on Prolific of 80–100%. We further prescreened for participants' highest education level earned to determine current and recent undergraduate status, as well as highest parental education level to determine college generation status. We excluded one participant for failing to pass an attention check, bringing the final sample to 399. Table 1 details demographic information focusing specifically on the 391 participants for whom we were able to determine college generation status. Analyses not including college generation status included data from the entire set of 399 participants.

Table 1. Participant demographic information disaggregated by college generation status.

Demographic Category	First-Gen (n = 169)	Continuing-Gen (n = 222)
Gender		
Man	45%	45%
Woman	50%	50%
Non-binary, non-conforming, bigender, transgender	5%	5%
Race/Ethnicity		
African American, Black, Caribbean	11%	9%
Asian, Asian American	18%	18%
Caucasian, White, European American	42%	62%
Latinx, Hispanic, Chicanx	20%	4%
Middle Eastern	1%	1%
Native American, American Indian, Alaska Native	1%	1%
Mixed Race/Ethnicity	8%	6%
Prefer Not To Respond	0%	0%
Academic Level		
Current undergraduate students	48%	51%
Earned undergraduate degree	51%	48%
Unknown	2%	1%

We chose to include recent graduates, and not just current undergraduate students, because it allowed us to recruit from a larger and more diverse population. Moreover, most studies to date have not examined these effects in recent graduates, and there is good reason to think that cultural mismatch experiences extend post-graduation, especially within the context of U.S. professional workplaces (e.g., Stephens et al., 2019).

2.2. Procedure, Materials, and Measures

Participants were randomly assigned to and completed one of four value affirmation conditions: independent (n = 102), interdependent (n = 105), combined (n = 115), and control (n = 77). It is unclear why fewer participants were run in the control condition than in the other conditions, but it may have been due to an issue with recruiting in Prolific. Importantly, however, participants were randomly assigned to the conditions. We thus have no reason to think that the participants in the four conditions had pre-existing differences. After completing the value affirmation exercise (or not), all participants read an educational passage at their own pace and were then immediately tested on their memory of the contents of that passage. Finally, all participants completed two rating scales, were debriefed on the study's purpose, and were compensated USD 10 for participating. The entire study took about 30 min to complete per participant. All study procedures and materials were approved as exempt by the UC Santa Cruz Institutional Review Board #HS-FY2022-372.

2.2.1. Value Affirmation Manipulation

The value affirmation conditions were modeled after the designs employed by Tibbetts et al. (2016) and Hecht et al. (2021). Specifically, we selected three values associated with interdependence (i.e., Relationships with Friends and Family, Spiritual or Religious Values, and Belonging to a Social Group), three values associated with independence (i.e., Independence, Learning and Gaining Knowledge, and Curiosity), and two neutral values that were neither highly associated with interdependence nor independence (i.e., Being Good at Art, Government and Politics). The interdependent and independent values have been used in past research, and past participants have been shown to commonly select and write about them (Harackiewicz et al., 2016). The neutral values have also been used in past research, but participants have tended to select them least. The neutral values were included as alternatives to create an illusion of choice while still encouraging participants to write about independence and interdependence (Tibbetts et al., 2016). The value affirmation prompts are available on the Open Science Framework (OSF).

Independent condition: Participants were presented with five values (three related to independence: Independence, Learning and Gaining Knowledge, Curiosity; two neutral: Being Good at Art, Government and Politics) and asked to select the two values that were most important to them. Once the two values were selected, participants were asked to write two paragraphs (4–5 sentences each) explaining why each value was important to them.

Interdependent condition: Participants were presented with five values (three related to interdependence: Relationships with Friends and Family, Spiritual or Religious Values, Belonging to a Social Group; two neutral: Being Good at Art, Government and Politics) and asked to select the two values that were the most important to them. Once the two values were selected, participants were asked to write two paragraphs (4–5 sentences each) explaining why each value was important to them.

Combined condition: Participants were presented with four values (two related to independence: Learning and Gaining Knowledge, Curiosity; two related to interdependence:

Relationships with Friends and Family, Belonging to a Social Group) and asked to select the two values that were most important to them. More specifically, they were asked to select one value related to independence, and one value related to interdependence. Neutral values were excluded from this condition to ensure that participants reflected on both independence and interdependence. The instructions for writing two paragraphs mirrored the other conditions.

Control condition: Participants did not engage in any form of value affirmation exercise. Although prior research has sometimes employed a control condition in which participants write about values that are not important to them but are important to others, such an activity could inadvertently act as a kind of value affirmation exercise, especially if participants were prompted to reflect on values important to their family members. Thus, we asked participants to complete an anagram task instead. Participants rearranged nine scrambled letters into as many meaningful 4 to 6 letter words as possible, with 44 words to find. Participants identified, on average, less than a fifth (M = 0.16, SD = 0.10) of all the words possible, and did not receive feedback.

Two independent raters scored each value affirmation response for the absence (scored as 0) or presence (scored as 1) of independent and interdependent themes (see Hecht et al., 2021, Appendix B for scoring criteria). To be coded for independence, at least one of the following criteria needed to be met: expressing the value of independence for the self, valuing an activity because it is done alone, and showing that the participant values their own autonomy. To be scored as interdependent, the criteria included valuing an activity because it is done with others, feeling like part of a group of people because of a certain value or while engaging in a certain activity, and mentioning thoughts of interdependence. A given response could receive scores for both independence and interdependence. Our inter-rater reliability on coding was near-perfect for independence (Cohen's kappa = 0.99; Landis & Koch, 1977), and perfect for interdependence (Cohen's kappa = 1.00).

Through this analysis, we were able to confirm that participants expressed values generally as assigned by their condition. Participants in the independent condition were much more likely to express values related to independence (M = 0.96) than interdependence (M = 0.08), and participants in the interdependent condition were much more likely to express values related to interdependence (M = 0.82) than independence (M = 0.29). Participants in the combined condition showed a slight bias towards expressing values related to independence (M = 0.60) over independence (M = 0.45).

A total of 18 participants (2, 2, and 14 in the independent, interdependent, and combined conditions, respectively) did not follow the instructions exactly as directed, either by writing paragraphs that appeared to be off-topic or by clearly focusing on values other than those to which they were assigned. This was particularly common in the combined condition, with participants continuing to focus on the first value in their second paragraph instead of switching to the second value. For the purposes of ensuring internal validity, we opted to include these participants in their assigned conditions (i.e., reassigning participants to other conditions could create selection effects, thus leading to pre-existing differences between groups in the four conditions). Moreover, all participants experienced the affirmation intervention as assigned, and so any variance in how the participants responded to the intervention was important to include when considering potential applied implications.

2.2.2. Learning and Recalling Information from an Educational Passage

To examine the potential effects of the value affirmation exercises on learning, we used an educational passage about the science of ground water. The passage was 856 words

long and taken from the U.S. Geological Survey website (Clark & Briar, 1993). Participants were instructed to read the passage at their own pace and were told that they would be tested on their ability to recall information from the passage. They were also told that the passage was unrelated to the exercises completed earlier. After reading the passage, the participants were immediately asked to type out as much information as they could recall from the passage in any order and not to worry about spelling or grammar. Participants were given as much time as they needed to complete the recall task, opting to end the task when they felt they could not recall any additional information.

For the purpose of scoring recall performance, a total of 52 idea units were identified in the passage. Two raters read each participant's recall output and determined whether each idea unit was addressed in their output. Credit was given for idea units even if they were only partially addressed, regardless of word order, spelling, or grammar. Credit was not given for idea units that were either not addressed or that were addressed incorrectly. To calculate recall performance, the number of retrieved idea units was divided by 52 for each participant. Inter-rater reliability was perfect between the two raters (intraclass correlation coefficient = 1.00). Any disagreements in the scoring of individual items were resolved by a third rater prior to analysis.

2.2.3. Additional Measures

Two measures were administered after completing the free-recall task: the Cultural Congruity Scale (Gloria & Kurpius, 1996) and the Independent and Interdependent Motivations Scale (Stephens et al., 2012a). We hoped to use the responses on these scales to better understand how individual differences in cultural congruity and motivations to attend college might affect the impact of value affirmation exercises. Upon further reflection, however, we opted to focus on independent and interdependent motivations. Although cultural congruity can be useful to measure in many contexts, it is not strongly aligned with how Stephens et al. (2012a) define cultural mismatch (i.e., the clash between independence and interdependence) and, thus, may tap into something different than what we had hoped to measure.

The Independent and Interdependent Motivations Scale includes 12 items that assess the extent to which participants endorse independent or interdependent motives for attending college from 1 (strongly disagree) to 6 (strongly agree) (Stephens et al., 2012a). The scale consists of six independent items (e.g., Expand my knowledge of the world) and six interdependent items (e.g., Help my family out after I'm done with college). The items were averaged to create two subscales, with higher scores indicating higher independent (Cronbach's alpha = 0.87, *Mean* = 5.08, *SD* = 0.74) and interdependent (Cronbach's alpha = 0.82, *Mean* = 3.98, *SD* = 1.10) motivations.

3. Results

We first examined recall performance on the ground water passage as a function of college generation status and value affirmation condition. Specifically, we employed a 2 (first-generation vs. continuing-generation) × 4 (independent vs. interdependent vs. combined vs. control) between-subjects ANOVA, with the proportion of idea units (out of 52) included as the dependent variable. Neither of the main effects were statistically significant. Specifically, recall failed to differ significantly as a function of college generation status (first-generation: M = 29.5%, SE = 1.1%; continuing-generation: M = 28.5%, SE = 1.0%)—$F(1, 383) = 0.46$, $MSE = 0.02$, $p = 0.50$, $\eta p^2 = 0.00$—or Value Affirmation Condition (independent: M = 30.7%, SE = 1.5%; interdependent: M = 27.6%, SE = 1.4%; combined: M = 29.1%, SE = 1.4%; control: M = 28.6%, SE = 1.7%)—$F(3, 383) = 0.82$, $MSE = 0.02$, $p = 0.48$, $\eta p^2 = 0.01$.

As shown in Figure 1, we also failed to find evidence of a significant interaction between the two factors: $F(3, 383) = 0.54$, $MSE = 0.02$, $p = 0.66$, $\eta p^2 = 0.00$. Recall performance was approximately the same across value affirmation conditions for both the first-generation participants (independent: $M = 31.1\%$, $SE = 2.2\%$; interdependent: $M = 28.1\%$, $SE = 2.2\%$; combined: $M = 28.3\%$, $SE = 2.0\%$; control: $M = 30.5\%$, $SE = 2.6\%$) and continuing-generation participants (independent: $M = 30.3\%$, $SE = 1.9\%$; interdependent: $M = 27.1\%$, $SE = 1.9\%$; combined: $M = 29.9\%$, $SE = 1.8\%$; control: $M = 26.7\%$, $SE = 2.2\%$). Including age, gender, ethnicity, and educational status as covariates did not change the observed pattern of results.

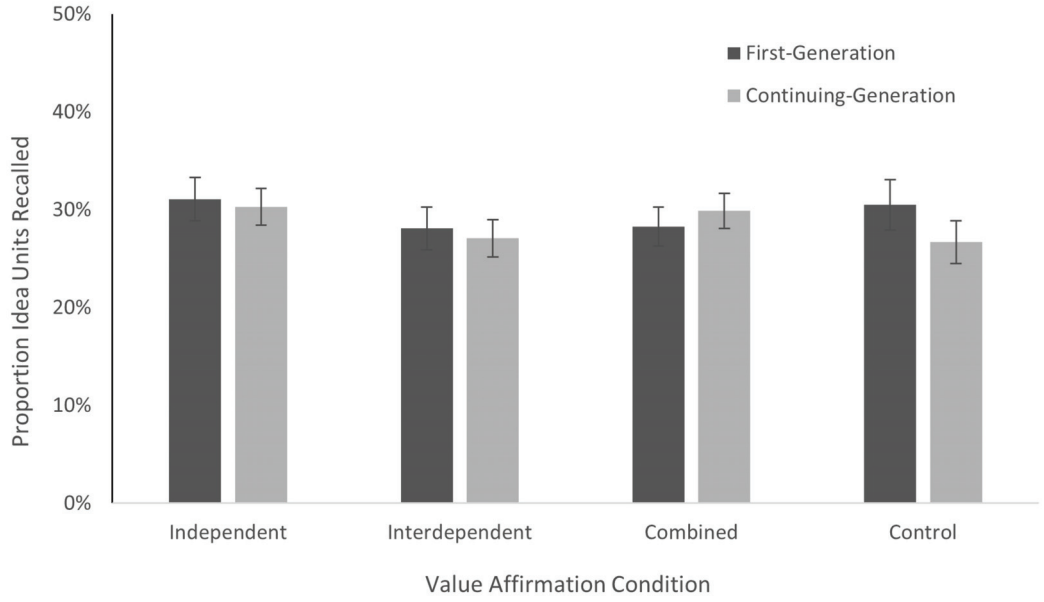


Figure 1. Value affirmation conditions by college generation status on units recalled. Note. Recall performance is shown for the four conditions as a function of college generation status. The error bars show the standard errors of the mean.

We conducted several follow-up analyses to confirm the robustness of these results. None of these analyses were pre-registered. First, we ran a one-way ANOVA including all participants (not just those for whom we were able to determine college generation status). The results were the same, with no evidence of a significant effect of Value Affirmation Condition: $F(3, 395) = 0.85$, $MSE = 0.02$, $p = 0.47$, $\eta p^2 = 0.01$. Then, we re-ran the 2×4 ANOVA reported above while excluding the 18 participants who did not complete the value affirmation task as instructed. These participants either wrote essays that were off-topic or did not focus on the particular values assigned. The pattern of results was once again the same, with no evidence of a main effect of college generation status—$F(1, 365) = 0.56$, $MSE = 0.02$, $p = 0.45$, $\eta p^2 = 0.00$—or value affirmation condition—$F(3, 365) = 1.02$, $MSE = 0.02$, $p = 0.39$, $\eta p^2 = 0.01$—or an interaction between the two factors—$F(3, 365) = 0.39$, $MSE = 0.02$, $p = 0.76$, $\eta p^2 = 0.00$.

Finally, to provide a more sensitive measure of the potential benefits of engaging in any value affirmation exercise, we collapsed all the participants who passed the attention check across the three value affirmation conditions into a single category. A 2 (first-generation vs. continuing-generation) $\times$ 2 (value affirmation vs. control) ANOVA once again failed to find any significant main effect of college generation status—$F(1, 387) = 1.04$, $MSE = 0.02$, $p = 0.31$, $\eta p^2 = 0.00$—or Value Affirmation Condition—$F(1, 387) = 0.06$, $MSE = 0.02$, $p = 0.80$, $\eta p^2 = 0.00$—or an interaction between the two factors—$F(1, 387) = 0.104$, $MSE = 0.02$, $p = 0.31$, $\eta p^2 = 0.00$. Indeed, neither first-generation participants (value affirmation: $M = 29.1\%$, $SE = 1.2\%$; control: $M = 30.5\%$, $SE = 2.6\%$; $t(167) = -0.52$, $p = 0.61$,

$d = -0.10$, 95% CI = [−0.49, 0.29]) nor continuing-generation participants (value affirmation: $M = 29.1\%$, $SE = 1.1\%$; control: $M = 26.7\%$, $SE = 2.2\%$; $t(220) = 0.96$, $p = 0.34$, $d = 0.16$, 95% CI = [−0.17, 0.50]) recalled significantly more idea units in the collapsed value affirmation conditions than they did in the control condition.

The results reported above failed to provide statistically significant evidence of the value affirmation exercises impacting recall performance. To better quantify evidence for the null, three Bayesian independent-samples *t*-tests were performed, comparing each of the value affirmation conditions with the control condition. All three Bayes Factors suggested moderate evidence for the null over the alternative (independent: BF = 5.8; interdependent: BF = 7.1; combined: BF = 8.1). A fourth analysis comparing the combination of the three affirmation conditions with the control condition also provided moderate evidence for the null, BF = 9.7. To give some context, a BF of 9.7 suggests that the data are approximately 9.7 times more likely under the null hypothesis than under the alternative hypothesis.

It is important to note that we failed to observe any significant differences in independent motivations (first-generation: $M = 5.11$, $SE = 0.06$; continuing-generation: $M = 5.06$, $SE = 0.05$) and interdependent motivations (first-generation: $M = 4.07$, $SE = 0.09$; continuing-generation: $M = 3.91$, $SE = 0.07$) as a function of generation status. Thus, to try to more directly examine the effect of motivations, we split the participants into two groups based on their responses to the Independent and Interdependent Motivations Scale. Specifically, we subtracted participants' independent motives score from their interdependent motives score to calculate a difference score for each participant and then rank-ordered the difference scores. A median split was used to create two groups of participants: one with relatively high interdependent motivations (independent motivations: $M = 4.86$, $SD = 0.83$; interdependent motivations: $M = 4.75$, $SD = 0.75$), and one with relatively high independent motivations (independent motivations: $M = 5.29$, $SD = 0.57$; interdependent motivations: $M = 3.26$, $SD = 0.87$). We then ran a 2 (interdependent motivations vs. independent motivations) × 4 (independent vs. interdependent vs. combined vs. control) between-subjects ANOVA, and as in the analysis reported above, no evidence of a significant interaction was observed: $F(3, 391) = 0.59$, $MSE = 0.02$, $p = 0.62$, $\eta p^2 = 0.01$. Recall performance was approximately the same across the value affirmation conditions both for participants with relatively high interdependent motivations (independent: $M = 28.4\%$, $SE = 2.2\%$; interdependent: $M = 26.8\%$, $SE = 2.0\%$; combined: $M = 26.6\%$, $SE = 1.9\%$; control: $M = 28.3\%$, $SE = 2.2\%$) and for participants with relatively high independent motivations (independent: $M = 32.2\%$, $SE = 1.9\%$; interdependent: $M = 28.81\%$, $SE = 2.0\%$; combined: $M = 32.2\%$, $SE = 2.2\%$; control: $M = 29.1\%$, $SE = 2.4\%$).

4. General Discussion

Given the persisting mismatch that many first-generation students navigate in university contexts, understanding which kinds of values to affirm offers a pathway for creating meaningful cultural alignments for students. We sought to investigate whether affirming for independent values, interdependent values, or both would affect the recall of an educational passage in current and recent first- and continuing-generation undergraduates relative to engaging in a non-affirmation control task. Contrary to past work documenting the separate benefits of independent (Miyake et al., 2010; Tibbetts et al., 2016; Yeager & Walton, 2011) and interdependent (Covarrubias et al., 2016; Hoshino-Browne et al., 2005) affirmations for marginalized students, we failed to find significant impacts of any of our affirmation conditions on recall performance.

Null results can be difficult to interpret. Yet they can also reveal important insights, especially regarding the sensitivity of replicating findings. We find value in pinpointing some of the differences in our study that might have produced different results from

earlier work. Indeed, though we closely followed intervention procedures outlined in previous studies (Covarrubias et al., 2016; Hecht et al., 2021), key design differences may partially explain our null effects. We detail these here and discuss possible directions for future research.

One important difference between the current study and most prior research on value affirmation is that the current study was conducted online. Dowell et al. (2021) also failed to observe significant benefits of value affirmation when data were collected online. Some online learning environments may not be conducive to developing particular skills, like reading comprehension, to the same degree as in-person learning (Patronis & Dubai, 2014). More importantly, because many participants likely completed the task at home and away from the university, they may have come into the experiment experiencing relatively less cultural mismatch than they would have in an in-person lab or classroom environment, thus diminishing the potential benefits of the value affirmation intervention. A more sensitive approach may be to focus on contexts in which learners are experiencing greater cultural mismatch to begin with. Still, given that undergraduates are increasingly learning in online environments, which has considerations both for mitigating and exacerbating cultural mismatches (Basch et al., 2022), our design becomes especially well-suited for examining online environments as critical aspects of the student experience. Future research might more directly test differences in performance tasks, including but not limited to the recall of to-be-learned information, across online versus in-person study designs.

Moreover, unlike other affirmation studies that have focused on undergraduate students still enrolled in university settings (Dowell et al., 2021; Hecht et al., 2021; Tibbetts et al., 2016), we included a diverse representation of both current undergraduate students and those who completed their degrees. Though they were diverse in terms of education level, this broader focus, especially by accessing participants with the resources to complete an online study, might also explain why our sample skewed toward more middle-class participants (with 79% reporting earning >USD 60,000 per year). Participants from middle-class backgrounds experience a different cultural mismatch than those from lower-income backgrounds. Moreover, students who have already completed their degree may also experience a different kind of mismatch or identity threat after graduating, especially when encountering the kind of learning task we administered in the current study. In this way, our decision to include college graduates could have contributed to the null effects that were observed. Future investigations with more analytic power might consider how educational status and/or income background inform experiences of a cultural mismatch and, thus, the impacts of the intervention. In fact, an even more direct test of this might include measuring participants' experiences of cultural mismatch that align with Stephens et al. (2012a)'s conceptualization and how an affirmation treatment speaks to those experiences. Research has offered creative ways of assessing more conceptually aligned and nuanced experiences of cultural mismatch (see Yılmaz et al., 2024), but has yet to assess the link to affirmation strategies.

Another possible interpretation of the current results is that, although the value affirmation exercises were effective in reducing cultural mismatch, the cognitive benefits accrued were not sufficient to enhance performance on the learning task. The cognitive processes involved in studying and recalling information from an educational passage may differ from those operating in prior affirmation tasks, like completing math problems (Covarrubias et al., 2016; Hecht et al., 2021; Shnabel et al., 2013; Tibbetts et al., 2016). Tasks involving processes like multiplying may involve a stronger or more direct working memory component than reading and recalling information and be relatively more likely to elicit identity threat. Thus, value affirmation exercises that enhance working memory may affect the former relatively more. If participants were given specific strategies that

depended on employing working memory in the learning task, perhaps a benefit would have been observed. However, the intervention may not have affected the spontaneous strategies participants employed, especially in an online learning environment.

Students often fail to use learning strategies known to boost long-term learning and transfer, choosing instead to engage in activities like rereading or highlighting. Techniques that optimize learning include self-testing, applying read material, and connecting to prior knowledge to practice what has been learned (Bjork & Bjork, 2011; Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007). Without direct instructions on these learning strategies and time to engage with them, values affirmations might not be as likely to impact learning. This is potentially the case when participants engage in reading material on topics they are less familiar with, like the science of ground water. Had participants been better able to connect the content to prior knowledge, especially via a relevant affirmation exercise, they could have created new retrieval pathways to support the learning process and potentially improve recall. Indeed, a subset of affirmation interventions include utility-value affirmations. Their main goal is to signal the usefulness or utility of the presented information, thus creating links to one's prior knowledge or future goals; these similarly boost performance outcomes for marginalized students (see Harackiewicz et al., 2016). Choosing a topic that better signals this utility might enhance the overall effectiveness of an affirmation, particularly when learning is involved. Another possibility is that one or more of the value affirmation exercises did have a positive impact on how participants studied the passage, but the immediate test was not sensitive to that impact. Work on desirable difficulties has shown that the benefits of many learning interventions can become greater over time (Bjork & Bjork, 2011). Thus, value affirmation exercises might be more likely to enhance performance on a delayed final test than on an immediate final test.

Reporting research regardless of whether significant results are obtained, or prior patterns are replicated, is critically important. Doing so provides a clearer picture of the theoretical mechanisms and boundary conditions of a given phenomenon. The current study was well-powered, with a final sample of 399 participants, while including multiple value affirmation exercises. Although reasons may exist to discount the current null results, the null results should also not be ignored. Future researchers should consider them carefully if attempting to explain when and how value affirmation exercises have the potential to promote learning and academic performance.

Although the current study failed to yield statistically significant results, the benefits of value affirmations in other settings are evident. Investigating the unique effects of independent, interdependent, and combined affirmations under different conditions provides important insights for sharpening their relevancy for students, especially those experiencing strong mismatches within the university. With a more deliberate design, that future work should necessarily explore the impact for something as important and central as learning and recalling new information. Understanding how value affirmation interventions can facilitate learning, including unpacking the conditions and mechanisms by which this happens, has valuable research and teaching implications. Better understanding the mechanisms behind value affirmations' effects on learning will help to delineate the ways in which we can most effectively apply cognitive psychology principles to support student learning. Such understandings will not only sharpen our knowledge about the range of impacts of affirmations but, if effective, will also serve to strengthen the tools we can employ in the classroom to reduce cultural mismatches and improve student learning.

Author Contributions: Individual author contributions are as follows: conceptualization, K.A., R.C., and B.C.S.; formal analysis, K.A., R.C., and B.C.S.; investigation, K.A., R.C., and B.C.S.; data curation: K.A. and B.C.S.; writing—original draft preparation, K.A.; writing—review and editing, K.A., R.C., and B.C.S.; project administration, K.A., R.C., and B.C.S. All authors have read and agreed to the published version of the manuscript.

Funding: The first author was funded by the University of California President's Postdoctoral Fellowship Program office while conducting the research.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved as exempt by the Institutional Review Board of UC Santa Cruz (protocol code: HS-FY2022-372; date of approval: 23 August 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data collected and materials employed are available on the Open Science Framework (OSF) project, accessed on 20 February 2025: https://osf.io/zb7kc/?view_only=cd8c0bbd59234310833ac24f711f3789.

Acknowledgments: We gratefully acknowledge the study participants. We also thank the following individuals for collecting and analyzing data, troubleshooting for accessibility, as well as for preparing and maintaining study platform pages, stimuli, and references: Sreejita Ghose, Nicholas T. Hilberg, Emily J. Kamalnathan, Peyton E. Marshall, Abigail Shamelashvili, Calia P. Thompson, Jordan Vogel, and Alyson Yin. Special thanks to the University of California President's Postdoctoral Fellowship Program office for funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Agarwal, P. K., & Roediger, H. L., III. (2018). Lessons for learning: How cognitive psychology informs classroom practice. *Phi Delta Kappan*, *100*(4), 8–12. [CrossRef]

Basch, S., Covarrubias, R., & Wang, S.-h. (2022). Minoritized students' experiences with pandemic-era remote learning inform ways of expanding access. *Scholarship of Teaching and Learning in Psychology*, *10*(4), 564–577. [CrossRef]

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Borman, G. D., Choi, Y., & Hall, G. J. (2021). The impacts of a brief middle-school self-affirmation intervention help propel African American and Latino students through high school. *Journal of Educational Psychology*, *113*(3), 605–620. [CrossRef]

Clark, D. W., & Briar, D. W. (1993). *What is ground water?* Report No. 93-643. U.S. Dept. of the Interior, U.S. Geological Survey. Available online: https://pubs.usgs.gov/of/1993/ofr93-643/pdf/ofr93-643.pdf (accessed on 16 June 2022).

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*(5791), 1307–1310. [CrossRef]

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science*, *324*(5925), 400–403. [CrossRef]

Cortés Pascual, A., Moyano Muñoz, N., & Quilez Robres, A. (2019). The relationship between executive functions and academic performance in primary education: Review and meta-analysis. *Frontiers in Psychology*, *10*, 1582. [CrossRef] [PubMed]

Covarrubias, R., Herrmann, S. D., & Fryberg, S. A. (2016). Affirming the interdependent self: Implications for latino student performance. *Basic and Applied Social Psychology*, *38*(1), 47–57. [CrossRef]

Covarrubias, R., Valle, I., Laiduc, G., & Azmitia, M. (2019). "You never become fully independent": Family roles and independence in first-generation college students. *Journal of Adolescent Research*, *34*(4), 381–410. [CrossRef]

Dowell, N. M., McKay, T. A., & Perrett, G. (2021). It's not that you said it, it's how you said it: Exploring the linguistic mechanisms underlying values affirmation interventions at scale. *AERA Open*, *7*, 1–19. [CrossRef]

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58. [CrossRef]

Gloria, A. M., & Kurpius, S. E. R. (1996). The validation of the cultural congruity scale and the university environment scale with Chicano/a students. *Hispanic Journal of Behavioral Sciences*, *18*(4), 533–549. [CrossRef]

Goudeau, S., Stephens, N. M., Markus, H. R., Darnon, C., Croizet, J.-C., & Cimpian, A. (2025). What causes social class disparities in education? The role of the mismatches between academic contexts and working-class socialization contexts and how the effects of these mismatches are explained. *Psychological Review*, *132*(2), 380–403. [CrossRef] [PubMed]

Gutchess, A., & Rajaram, S. (2023). Consideration of culture in cognition: How we can enrich methodology and theory. *Psychonomic Bulletin & Review*, *30*(3), 914–931. [CrossRef]

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology*, *111*(5), 745. [CrossRef] [PubMed]

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin and& Review*, *19*, 126–134. [CrossRef]

Hecht, C. A., Priniski, S. J., Tibbetts, Y., & Harackiewicz, J. M. (2021). Affirming both independent and interdependent values improves achievement for all students and mitigates cultural mismatch for first-generation college students. *Journal of Social Issues*, *77*(3), 851–887. [CrossRef]

Hoshino-Browne, E., Zanna, A. S., Spencer, S. J., Zanna, M. P., Kitayama, S., & Lackenbauer, S. (2005). On the cultural guises of cognitive dissonance: The case of easterners and westerners. *Journal of Personality and Social Psychology*, *89*(3), 294. [CrossRef]

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224. [CrossRef]

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. [CrossRef]

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, *330*(6008), 1234–1237. [CrossRef]

Pan, S. C., & Bjork, R. A. (2024). Acquiring an accurate mental model of human learning: Toward an owner's manual. In M. J. Kahana, & A. D. Wagner (Eds.), *Oxford handbook of memory* (pp. 2005–2034). Oxford University Press. [CrossRef]

Patronis, M., & Dubai, A. (2014). The Effect of using the iPad on students' performance in writing and reading comprehension: Pilot Study Report. *Arab World English Journal*, 67–80. Available online: https://api.semanticscholar.org/CorpusID:212411785 (accessed on 16 June 2022).

Phillips, L. T., Stephens, N. M., Townsend, S. S., & Goudeau, S. (2020). Access is not enough: Cultural mismatch persists to limit first-generation students' opportunities for achievement throughout college. *Journal of Personality and Social Psychology*, *119*(5), 1112–1131. [CrossRef]

Rajaram, S., & Barber, S. (2024). Retrieval Processes in Memory. In H. L. Roediger III (Ed.), *Learning and Memory: A Comprehensive Reference* (Vol. 2, pp. 261–284). Elsevier.

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. *Psychological Review*, *115*(2), 336–356. [CrossRef] [PubMed]

Shnabel, N., Purdie-Vaughns, V., Cook, J. E., Garcia, J., & Cohen, G. L. (2013). Demystifying values-affirmation interventions: Writing about social belonging is a key to buffering against identity threat. *Personality and Social Psychology Bulletin*, *39*(5), 663–676. [CrossRef] [PubMed]

Steele, C. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 21). Academic Press. [CrossRef]

Stephens, N. M., Fryberg, S. A., Markus, H. R., Johnson, C. S., & Covarrubias, R. (2012a). Unseen disadvantage: How American universities' focus on independence undermines the academic performance of first-generation college students. *Journal of Personality and Social Psychology*, *102*(6), 1178. [CrossRef] [PubMed]

Stephens, N. M., Townsend, S. S., & Dittmann, A. G. (2019). Social-class disparities in higher education and professional workplaces: The role of cultural mismatch. *Current Directions in Psychological Science*, *28*(1), 67–73. [CrossRef]

Stephens, N. M., Townsend, S. S., Markus, H. R., & Phillips, L. T. (2012b). A cultural mismatch: Independent cultural norms produce greater increases in cortisol and more negative emotions among first-generation college students. *Journal of Experimental Social Psychology*, *48*(6), 1389–1393. [CrossRef]

Szmalec, A., Verbruggen, F., Vandierendonck, A., & Kemps, E. (2011). Control of interference during working memory updating. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 137–151. [CrossRef]

Tibbetts, Y., Harackiewicz, J. M., Canning, E. A., Boston, J. S., Priniski, S. J., & Hyde, J. S. (2016). Affirming independence: Exploring mechanisms underlying a values affirmation intervention for first-generation students. *Journal of Personality and Social Psychology*, *110*(5), 635–659. [CrossRef]

Tibbetts, Y., Priniski, S. J., Hecht, C. A., Borman, G. D., & Harackiewicz, J. M. (2018). Different institutions and different values: Exploring first-generation student fit at 2-year colleges. *Frontiers in Psychology*, *9*, 502. [CrossRef]

Valle, I., & Covarrubias, R. (2024). Stories of conflict, agency, and hope: Master narratives underlying home-school mismatches for first-generation Latina students. *Qualitative Psychology*, *11*(3), 465. [CrossRef]

Vasquez-Salgado, Y., Greenfield, P. M., & Burgos-Cienfuegos, R. (2015). Exploring home-school value conflicts: Implications for academic achievement and well-being among Latino first-generation college students. *Journal of Adolescent Research*, *30*(3), 271–305. [CrossRef]

Wang, Q. (2009). Are Asians forgetful? Perception, retention, and recall in episodic remembering. *Cognition*, *111*(1), 123–131. [CrossRef] [PubMed]

Wang, Q. (2021). The cultural foundation of human memory. *Annual Review of Psychology*, *72*(1), 151–179. [CrossRef] [PubMed]

Wu, Z., Spreckelsen, T. F., & Cohen, G. L. (2021). A meta-analysis of the effect of values affirmation on academic achievement. *Journal of Social Issues*, *77*(3), 702–750. [CrossRef]

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, *81*(2), 267–301. [CrossRef]

Yılmaz, E., Phalet, K., & De Leersnyder, J. (2024). Putting cultural mismatch theory to the test: Cultural fit of self-construal in predicting student outcomes. *Journal of Social Issues*, *80*(4), 1379–1407. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Behavioral Sciences Editorial Office
E-mail: behavsci@mdpi.com
www.mdpi.com/journal/behavsci



Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Guest Editors. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editors and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.





Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-8214-4