Journal of Intelligence

Special Issue Reprint

Psycho-Educational Assessments

Theory and Practice

Edited by
Okan Bulut

mdpi.com/journal/jintelligence

MDPI

Psycho-Educational Assessments: Theory and Practice

Psycho-Educational Assessments: Theory and Practice

Editor

Okan Bulut

MDPI

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

Editor
Okan Bulut
Centre for Research in Applied
Measurement and Evaluation
University of Alberta
Edmonton, AB Canada

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Journal of Intelligence* (ISSN 2079-3200) (available at: www.mdpi.com/journal/jintelligence/special_issues/Psycho-Educational_Assessments_Theory_Practice).

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

Lastname, Firstname, Firstname Lastname, and Firstname Lastname. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-0552-5 (Hbk)
ISBN 978-3-7258-0551-8 (PDF)
doi.org/10.3390/books978-3-7258-0551-8

Contents

About the Editor

Okan Bulut

Okan Bulut is an Associate Professor of the Measurement, Evaluation, and Data Science program and a researcher at the Centre for Research in Applied Measurement and Evaluation at the University of Alberta. His research interests lie at the intersection of artificial intelligence (AI), educational data mining and learning analytics. Through the utilization of AI-driven algorithms and natural language processing, he seeks to create intelligent systems that can dynamically adapt to learner preferences, cognitive styles and performance trajectories.

Preface

Psycho-educational assessments, such as intelligence tests, cognitive test batteries and behavioral measures, serve as invaluable tools for school psychologists and educators. They provide profound insights into children's learning and behavioral profiles, enhancing our understanding of their academic and cognitive capacities. This reprint of the Special Issue titled "Psycho-Educational Assessments: Theory and Practice" includes a collection of sixteen articles that delve into the different facets of psycho-educational assessments. Each article serves as a testament to our remarkable journey in understanding and enhancing the psycho-educational assessment landscape, from exploring the efficacy of psycho-educational assessments in diagnosing a range of learning difficulties to different statistical techniques for analyzing psycho-educational constructs, such as local structural equation models and multidimensional scaling. As the editor of this Special Issue, I anticipate that the articles showcased herein will serve as catalysts for meaningful discourse, igniting innovative ideas and enriching the ongoing evolution of psycho-educational assessment.

Okan Bulut
Editor

Journal of
Intelligence

MDPI

Editorial

Psycho-Educational Assessments: Theory and Practice

Okan Bulut

Centre for Research in Applied Measurement and Evaluation, Faculty of Education, University of Alberta, Edmonton, AB T6G 2G5, Canada; bulut@ualberta.ca

Psycho-educational assessments, such as intelligence tests, cognitive test batteries, and behavioral measures, serve as invaluable tools for school psychologists and educators. They provide profound insights into children's learning and behavioral profiles, enhancing our understanding of their academic and cognitive capacities. By employing these assessments, professionals can pinpoint each student's individual strengths and weaknesses. Moreover, psycho-educational assessments play a pivotal role in identifying various educational needs, including learning disabilities, intellectual differences, social-emotional challenges, and giftedness. These assessments offer a comprehensive view of students' abilities and potential hurdles they may encounter in their academic journey. Beyond diagnosis, the results of psycho-educational assessments can also inform the development of tailored interventions and support programs. They enable educators to implement timely strategies that address the unique requirements of students, fostering an inclusive learning environment where every child can thrive.

This Special Issue on "Psycho-Educational Assessments: Theory and Practice," presented by the Journal of Intelligence, provided us with the opportunity to curate a collection of sixteen articles that delve into different facets of psycho-educational assessments. Each article serves as a testament to our remarkable journey in understanding and enhancing the psycho-educational assessment landscape. For this Special Issue, researchers were asked to present findings of empirical or methodological research and theoretical work related to the design, use, analysis, interpretation, and reporting processes of psycho-educational assessments. All articles submitted for publication underwent a rigorous peer review based on the review standards established by the Journal of Intelligence.

For the sake of brevity, I want to touch upon the focal points of this Special Issue and highlight key articles. The authors dedicated their efforts to exploring the identification of cognitive, social, emotional, and behavioral challenges across diverse groups, ranging from children aged 6 to 11, individuals with autism spectrum disorder, and verbally gifted children to those from regions like China and the United Arab Emirates. These studies primarily relied on psycho-educational assessments, employing intelligence tests, behavioral checklists, and psychological scales as primary data collection tools. Notably, intelligence tests, given their increasing significance in educational contexts, featured prominently in this Special Issue. Several studies leveraged well-established cognitive assessment batteries, including the Wechsler Intelligence Scale for Children, the Wechsler Individual Achievement Test, and the Woodcock–Johnson Tests of Academic Achievement and Cognitive Abilities. In addition to empirical investigations, the Special Issue encompassed methodological inquiries. For instance, Gao et al. (2022) conducted an empirical study focusing on process data indicators to explore problem-solving styles within technology-rich environments. Furthermore, Meyer and Reynolds (2022) demonstrated a methodological approach, multidimensional scaling, to scrutinize correlations between cognitive abilities and academic achievement test scores.

As the editor of this special issue, I anticipate that the articles showcased herein will serve as catalysts for meaningful discourse, igniting innovative ideas and enriching the ongoing evolution of psycho-educational assessment. Moreover, these featured articles will

check for updates

Citation: Bulut, Okan. 2024. Psycho-Educational Assessments: Theory and Practice. *Journal of Intelligence* 12: 31. https://doi.org/10.3390/jintelligence12030031

Received: 15 February 2024
Accepted: 26 February 2024
Published: 5 March 2024

provide valuable insights into enhancing various aspects of the design, utilization, analysis, interpretation, and reporting processes involved in psycho-educational assessments. The contributions within this collection delve into diverse facets, from exploring the efficacy of psycho-educational assessments in diagnosing a range of learning difficulties to different statistical techniques for analyzing psycho-educational constructs, such as local structural equation models and multidimensional scaling.

In closing this editorial summary, I wish to express my heartfelt gratitude to the esteemed authors whose expertise and dedication have profoundly enriched this Special Issue. A special acknowledgment extends to all the diligent reviewers who generously shared their valuable insights, contributing significantly to the refinement of the submitted papers. Additionally, I express my sincere appreciation to the Editorial Team of the Journal of Intelligence for entrusting me with the privilege of curating this Special Issue. This Special Issue would not have been possible without their unwavering support and meticulous efforts.

Conflicts of Interest: The author declares no conflicts of interest.

References

Gao, Yizhu, Xiaoming Zhai, Okan Bulut, Ying Cui, and Xiaojian Sun. 2022. Examining Humans' Problem-Solving Styles in Technology-Rich Environments Using Log File Data. *Journal of Intelligence* 10: 38. [CrossRef] [PubMed]

Meyer, Em M., and Matthew R. Reynolds. 2022. Multidimensional Scaling of Cognitive Ability and Academic Achievement Scores. *Journal of Intelligence* 10: 117. [CrossRef] [PubMed]

Journal of
Intelligence

MDPI

Article

Multidimensional Scaling of Cognitive Ability and Academic Achievement Scores

Em M. Meyer [1,*] and Matthew R. Reynolds [2]

[1] Department of Counseling, School Psychology and Family Science, College of Education, University of Nebraska, Kearney, NE 68849, USA
[2] Department of Educational Psychology, School of Education and Human Sciences, University of Kansas, Lawrence, KS 66045, USA
* Correspondence: meyerem@unk.edu

Abstract: Multidimensional scaling (MDS) was used as an alternate multivariate procedure for investigating intelligence and academic achievement test score correlations. Correlation coefficients among Wechsler Intelligence Scale for Children, Fifth Edition (WISC-5) and Wechsler Individual Achievement Test, Third Edition (WIAT-III) validity sample scores and among Kaufman Assessment Battery for Children, Second Edition (KABC-II) and Kaufman Test of Educational Achievement, Second Edition (KTEA-2) co-norming sample scores were analyzed using multidimensional scaling (MDS). Three-dimensional MDS configurations were the best fit for interpretation in both datasets. Subtests were more clearly organized by CHC ability and academic domain instead of complexity. Auditory-linguistic, figural-visual, reading-writing, and quantitative-numeric regions were visible in all models. Results were mostly similar across different grade levels. Additional analysis with WISC-V and WIAT-III tests showed that content (verbal, numeric, figural) and response process facets (verbal, manual, paper-pencil) were also useful in explaining test locations. Two implications from this study are that caution may be needed when interpreting fluency scores across academic areas, and MDS provides more empirically based validity evidence regarding content and response mode 6processes.

Keywords: multidimensional scaling; intelligence; academic achievement; complexity

Citation: Meyer, Em M., and Matthew R. Reynolds. 2022. Multidimensional Scaling of Cognitive Ability and Academic Achievement Scores. *Journal of Intelligence* 10: 117. https://doi.org/10.3390/jintelligence10040117

Received: 29 September 2022
Accepted: 21 November 2022
Published: 1 December 2022

1. Introduction

Individually administered, norm-referenced intelligence and academic achievement tests are important components of comprehensive psychoeducational evaluations (Benson et al. 2019; Rabin et al. 2016). Scores from these tests provide evidence of expected learning (intelligence), demonstrated learning (academic achievement), and cognitive and academic strengths and weaknesses (Benson et al. 2019). Results from these tests are combined with other assessment information to make diagnoses, assist with educational eligibility decisions, and plan individualized educational programs and interventions (Lichtenberger and Breaux 2010; Mather and Abu-Hamour 2013). Psychologists depend upon up-to-date information to improve the likelihood of valid interpretations of their scores.

1.1. Intelligence and Academic Achievement Tests Are Multidimensional and Related

Individually administered intelligence tests measure general intelligence and more specific abilities such as problem-solving, auditory processing, and memory (Kaufman and Kaufman 2004; Wechsler 2014). The latent structure of intelligence is hierarchical and multidimensional (Caemmerer et al. 2020; Reynolds and Keith 2017), and it is organized as such in Cattell-Horn-Carroll (CHC) theory (McGrew 2009; Schneider and McGrew 2018). CHC abilities such as novel problem solving, auditory processing, and memory give rise to observable differences in intelligence test scores. These CHC abilities are interrelated and correlate with a general factor of intelligence, or *g*. Fluid reasoning (Gf) typically has

the strongest correlation with *g* (Carroll 1993; Jensen 1998). Other broad abilities include comprehension-knowledge (Gc), short-term working memory (Gsm or Gwm), long-term storage and retrieval (Glr), visual processing (Gv), and auditory processing (Ga).

Individually administered academic achievement tests measure acquired skills in academic domains such as reading, writing, and math. Within each domain, there are levels of complexity: basic skills (e.g., knowledge of letter-sound relations), fluency (e.g., automatic word reading), and higher-order thinking (e.g., reading comprehension). Composite achievement test scores can reflect performance within a domain or across domains according to complexity (Mather and Wendling 2015). Acquired knowledge, including the types of skill domains measured with academic achievement tests such as reading and writing or math, are also described in CHC theory (Schneider and McGrew 2018).

A valid interpretation of intelligence and academic achievement scores depends on understanding how and why the test scores correlate with each other. Most studies of intelligence and academic achievement test score correlational structures have used structural equation modeling, factor analysis, and regression analysis. Correlations between *g* and latent general academic achievement are very strong (Kaufman et al. 2012). Intelligence, however, contributes to specific academic achievement domains generally and via specific CHC abilities (Floyd et al. 2007; Gustafsson and Balke 1993; Niileksela and Reynolds 2014). For example, verbal comprehension (understanding words and their relations) and auditory processing (perception and manipulation of sound) affect reading in addition to the effects of *g* (Garcia and Stafford 2000; Keith 1999; Vanderwood et al. 2002).

CHC broad abilities influence both specific and broad areas of reading, math, and writing (Benson et al. 2016; Caemmerer et al. 2018; Cormier et al. 2016; Cormier et al. 2017; Floyd et al. 2003; Hajovsky et al. 2020; McGrew and Wendling 2010; Niileksela et al. 2016). As such, tests that draw upon CHC abilities relate to tests of specific academic achievement. In addition, intelligence and academic achievement tests may share characteristics that contribute to their correlations such as task complexity (basic skills, fluency, or higher-order thinking or problem-solving), stimuli (words, numbers, or pictures), and examinee response modes (oral response, manipulation of materials, or written). Empirical studies of the correlations between intelligence and achievement scores should try to incorporate these other shared characteristics.

1.2. Validity Evidence from Multidimensional Scaling

AERA, APA, and NCME Standards for Educational and Psychological Testing (American Educational Research Association et al. 2014) include test content and response processes as sources of validity evidence. These types of evidence, however, are rarely demonstrated with data, rather they are described by conducting an alignment study or expert panel review. Multidimensional scaling (MDS) is an unrestricted, multivariate technique for analyzing correlations and exploring test score interrelations in a visual way. It is also an empirical method used to evaluate validity evidence based on content (Li and Sireci 2013) and response process (Cohen et al. 2006).

MDS has been applied to intelligence test scores alone (Cohen et al. 2006; Guttman and Levy 1991; McGill 2020; Meyer and Reynolds 2018; Tucker-Drob and Salthouse 2009) and together with academic achievement scores (Marshalek et al. 1983; McGrew 2012; McGrew et al. 2014; Snow et al. 1984). However, in the context of the thousands of factor-analytic studies used with intelligence and academic achievement scores, MDS has been applied rarely.

MDS puts all of the variables (e.g., tests) in continuous, geometric space based on their intercorrelations, and MDS "maps" of variables are created (Borg and Groenen 2005; Tucker-Drob and Salthouse 2009). Highly related scores are spatially closer in MDS maps, facilitating the interpretation of shared characteristics. For example, in previous MDS research with intelligence test constructs, verbal comprehension tests clustered together in one area of the map that was separate from other clusters of tests (Cohen et al. 2006; Marshalek et al. 1983; Meyer and Reynolds 2018).

One advantage of MDS is that it allows shared test characteristics to emerge because the tests are displayed in continuous space, and all variables remain in the model for interpretation instead of being reduced to a smaller number of variables as in principal components analysis or factor analysis. Objects in the MDS map can "differ along many dimensions simultaneously" (Snow et al. 1984, p. 89). Another advantage of MDS is that it does not impose as many expectations as in factor analysis (Tucker-Drob and Salthouse 2009).

1.3. MDS with Intelligence and Academic Achievement

Intelligence and academic achievement test score correlations can be analyzed together with MDS procedures. Pairs of tests with higher correlations are expected to be closer to each other in the MDS configuration. Additionally, the center of the map is the shortest distance from all other points, so tests with the strongest correlations with all other tests are expected to be in the center of the MDS map. Tests that do not correlate highly with all other tests are expected to be farther from the center of the MDS configuration, typically in clusters of highly related tests (Marshalek et al. 1983; McGrew et al. 2014). Although historically the focus has been on test characteristics and not latent abilities, conceptually, tests in the center are at-times more "*g*-related" and tests are often grouped by CHC broad abilities (Meyer and Reynolds 2018). Therefore, studies have indicated that both test complexity and test content are mapped, but the interpretations often parallel those from factor analysis and CHC theory (Marshalek et al. 1983; Guttman and Levy 1991).

Shared test content (e.g., verbal tests) or CHC abilities are often useful in describing the organization of tests in clusters or regions of an MDS configuration (Marshalek et al. 1983; McGrew et al. 2014; Meyer and Reynolds 2018). For example, fluid reasoning tests may be clustered together, comprehension-knowledge tests clustered together, and visual processing tests clustered together. Within those CHC clusters, however, tests with higher *g*-loadings are often located closer to the center of the map forming what is called a "radex".

Due to this arrangement in space, researchers have also interpreted test complexity as another dimension observable from MDS analysis of intelligence test scores. A typical pattern of MDS analysis of intelligence tests alone is for the most complex tests (higher-order thinking) to be near the center of the map (Marshalek et al. 1983; Tucker-Drob and Salthouse 2009), even though they differ in content. Tests near the center also often have the highest *g*-loadings, although this is not always the case (e.g., McGrew et al. 2014). For example, Marshalek et al. (1983) found fluid reasoning tests were closest to the center of the map–and these tests were considered the most cognitively complex tests. Comprehension-knowledge and visual processing tests were in the intermediate range of complexity. Memory and speed tests were farthest from the center of the MDS map. Therefore, the map seemed to place complex tests in the center. They described the continuum radiating out from the center as going from complex-general to simple-specific. Notably, however, the arrangement of the types of tests that radiate from the center also appear to be associated with the magnitudes of correlations between the latent CHC abilities and the *g* factor-fluid reasoning is the strongest, followed by comprehension-knowledge, visual processing, memory, and then processing speed (Carroll 1993). Likewise, intelligence test scores are often interpreted via CHC theory, with a focus more on composites such as the IQ as an indicator of *g* and broad indexes as indicators of CHC broad abilities (e.g., top down).

Academic achievement test data, on the other hand, are often interpreted more from the bottom up, and based on shared content. For example, rather than general reading, the more basic task of word reading or nonword reading scores are interpreted first before considering general reading ability, which is to some extent dependent on basic reading skills (and more of an emergent construct). Academic achievement test data also provide a clearer example of how both test content and complexity dimensions may appear in an MDS configuration. For example, reading tests should be located separately from other academic achievement domains (e.g., mathematics) and radiate from the center in a straight line. Reading comprehension (a complex reading task) should be closest to the

center, and word recognition (the least complex reading task) should be farthest from the center. Reading fluency (intermediate complexity) would be located in the middle along an imaginary straight line connecting the reading comprehension and word recognition tests.

Most MDS research has been conducted with intelligence test scores. We are not aware of MDS studies with achievement tests only. MDS research with intelligence and academic achievement test scores analyzed together is limited to a few studies conducted 35 years ago, analysis of Woodcock-Johnson Psycho-Educational Battery—Revised and Woodcock-Johnson III data (McGrew 2012), and research reported in the Woodcock-Johnson IV test manual (McGrew et al. 2014).

McGrew et al. (2014) divided the Woodcock-Johnson IV Tests of Cognitive Abilities, Tests of Academic Achievement, and Tests of Oral Language normative data into age-based subsamples for MDS analysis. At least six notable general findings emerged from the analysis. First, CHC abilities described test location better than test content. For example, the Pair Cancellation test, a processing speed test that contains visual stimuli, was closer to processing speed tests than it was to visual processing tests. Second, "regions" emerged that often consisted of two or more CHC abilities: auditory-linguistic, figural-visual, reading-writing, quantitative-numeric, and speed-fluency regions. Shared components in these regions better explained test location than content—for example, Writing Fluency, Word Reading Fluency, and Math Facts Fluency tests clustered within the speed-fluency region, next to the reading-writing and quantitative-math regions. The shared speed component better explained the location of fluency tests not their content. Third, academic achievement tests were mostly separate from intelligence tests. In some instances with the achievement tests, tests were organized with higher-order thinking tests located closer to the center of the MDS map (e.g., Applied Problems math test) and simpler tests farther from the center (e.g., Calculation basic math test). Fourth, achievement tests clustered closer to some cognitive test clusters than others. Reading and writing tests were closer to comprehension-knowledge and auditory processing CHC clusters (in the auditory-linguistic region), and math tests were closer to working memory and visual processing CHC clusters (in the figural-visual region). Fifth, tests did not all radiate outward from complex to simple in exact order by complexity as indicated by *g*-loadings (e.g., McGrew et al. 2014). Sixth, there were slight changes across age groups. A general memory region was found in the 6–8 age group, but not in others—theoretically, because the relations among cognitive abilities and achievement constructs do change with age, some age-related changes may be expected (e.g., Hajovsky et al. 2014). In light of these findings, MDS research with measures other than the Woodcock-Johnson tests is needed. Were they findings specific to the Woodcock tests? Or are these more generalizable findings?

1.4. Facet Theory

MDS provides an opportunity to interpret test score relations in content (Guttman's mode of communication) and complexity (Guttman's rule inference) simultaneously. MDS is often associated; however, with Guttman's facet theory that organizes and defines observations, such as those elicited during intelligence assessment. His facet theory also included response mode (Guttman's mode of expression), which refers to how examinees respond to test items: oral, manual manipulation of materials (pegs, tiles, or blocks), and written (Guttman and Levy 1991). Response mode has been a useful explanation to test scores when three dimensional MDS maps are formed with intelligence scores (Cohen et al. 2006). Response processes are also considered important when evaluating validity evidence (American Educational Research Association et al. 2014), but rarely submitted to empirical analysis. Here, we wanted to consider response processes in our analysis. Hence, test scores may be correlated due to similar complexity (or how closely they related to psychometric g), content (or latent CHC broad abilities), and response processes. These three dimensions have emerged from a combination of Guttman's facet theory, CHC theory, and research using MDS to map intelligence score correlations (Cohen et al. 2006; Marshalek

et al. 1983). We wanted to investigate if they emerged in some way when including data from intelligence and achievement tests.

1.5. Purpose of the Study

The purpose of this study was to use MDS to analyze correlations among Wechsler cognitive and achievement tests and among Kaufman cognitive and achievement tests to better understand the relations among the scores. We did so for several reasons.

First, we wanted to use an alternative multivariate method to analyze intelligence and academic achievement test score correlations in combination. The majority of research has used factor analysis or some other form of structural equation modeling. Although MDS is used rarely, according to Snow et al. (1984, p. 88), because MDS "stays close to the original measures, requires minimal assumptions, and provides a simple representation, it is the method of choice when only one method is used." Limiting theory and findings to certain statistical methods, such as factor analysis, may obscure nuance and limit new and important findings, especially when assumptions and choices used in new research depend on previous findings based on the same method. Do important findings emerge from analyzing these data with MDS? Even if the analysis produces results that are similar to those from factor analysis, similar findings with alternative methods only bolster confidence in the previous findings.

Second, relatively few studies have used MDS to analyze intelligence test score correlations. Fewer have used MDS to analyze intelligence and academic achievement score correlations together. Explanations of findings from these studies have paralleled those from hierarchical factor analysis (Marshalek et al. 1983; Meyer and Reynolds 2018), such that CHC theory rather than test content may be used to explain clusters of tests in MDS space and how tests or clusters of tests radiate out from the center of MDS space (e.g., *g*-loadings are associated with a complexity dimension [fluid reasoning tests in the center of the map]). Those studies focused on intelligence test data, however. What happens when achievement tests are included? For example, does CHC theory still help to interpret findings? Or does test content also need to be considered.

Third, response processes are important aspects of test score validity. They are rarely evaluated using empirical methods, however. In addition to complexity and content (or their CHC parallels), do response processes help to understand correlations among intelligence and achievement tests when they are placed in multivariate space? If so, it may help with test score interpretation.

Fourth, revisions of popular tests, such as the Wechsler and Kaufman tests (Benson et al. 2019), have not been included in MDS studies with intelligence and achievement scores at all. Are findings related to the Woodcock tests generalizable? Specifically, McGrew et al.' (2014) analysis also suggested larger auditory-linguistic, figural-visual, reading-writing, quantitative-numeric, and speed-fluency conceptual regions are useful in describing the scores in space. We wanted to test the reliability and viability of those regions with data from different test batteries. They are useful conceptual categories, but they would be much less interesting and useful if they do not emerge in data outside of the Woodcock tests.

Last, we wanted to analyze whether these findings change across different developmental levels. On the one hand, analyzing data across ages or grades tests the reliability of the findings. Findings are not expected to drastically change across developmental levels so the findings should generally be consistent across the ages (Reynolds et al. 2007; Reynolds and Keith 2017). On the other hand, McGrew et al. (2014) found slight developmental changes and other research shows the relations between cognitive and academic may change with age—for example, comprehension knowledge is more highly associated with reading comprehension in adolescents than it is in younger children (Hajovsky et al. 2014). Therefore, because the Kaufman data were from a large sample, we divided the data into different grade level groups for analysis.To achieve our purpose, we asked the following research questions. Each question is accompanied by initial hypotheses. Each question applies to the different grade groups.

- Are complex tests in the center of the MDS configuration with less complex tests farther from the center of the MDS configuration?
- Intelligence and academic achievement tests of higher complexity were predicted to be near the center of the configuration and tests of lower complexity were predicted to be on the periphery (Marshalek et al. 1983). However, tests were not necessarily expected to all radiate outward from complex to simple tests in exact order by complexity as indicated by *g*-loadings (e.g., McGrew et al. 2014). Are intelligence tests and academic achievement tests clustered by CHC ability and academic content, respectively?
- Ga, Gc, Gv, Gf, and Gsm or Gwm tests were expected to cluster by CHC ability, and reading, writing, math, and oral language tests were expected to cluster by academic achievement area. Certain regions of academic achievement tests were predicted to align more closely with CHC ability factors. Reading and writing tests were predicted to be close to the Gc, Ga, and oral language tests. Math tests were predicted to be closer to the Gsm or Gwm, Gv, and Gf clusters.
- Are tests organized into auditory-linguistic, figural-visual, reading-writing, quantitative-numeric, and speed-fluency regions?
- Auditory-linguistic, figural-visual, reading-writing, quantitative-numeric, and speed-fluency regions were investigated in this study (McGrew et al. 2014). Gc tests, Ga tests, and oral language tests were predicted to cluster together with each other within an auditory-linguistic region. Reading and writing tests were predicted to be located in a reading-writing region. Gf tests were predicted to be in figural-visual or quantitative-numeric regions. Gv tests were predicted to be in a figural-visual region. Gsm or Gwm tests were predicted to be in the region that corresponded to the figural or numeric content (i.e., tests with pictures in the figural-visual region and tests with numbers in the quantitative-numeric region). Glr tests were not expected to be in just one region or in the same region of every configuration (McGrew et al. 2014).

2. Materials and Methods

2.1. Participants

2.1.1. Wechsler Sample Participants

Data for the MDS with the WISC-V and WIAT-III were correlations derived from participant scores in a WISC-V validity study (see Wechsler 2014) with an average testing interval a little over two weeks (M = 15.5, SD = 14.37). Data from 181 English-speaking children and adolescents between the ages of 6 and 16 were used. Demographics are in Table 1. Demographics were mostly similar to the WISC-V norming sample and therefore similar to the U.S. population in 2012.

Table 1. Demographic Information: WISC-V and WIAT-III Validity Sample, WISC-V Norming Sample.

Demographic Variable	% of Validity Sample *N* = 181
Sex	
Female	44.8
Male	55.2
Race/Ethnicity	
Asian	1.7
Black	19.9
Hispanic	21.0
Other	7.2
White	50.3
Highest Parental Education	
Grade 8 or less	2.2
Grade 9–12, no diploma	8.3
Graduated high school or GED	24.9
Some College/Associate Degree	35.4
Undergraduate, Graduate, or Professional degree	29.3

2.1.2. Kaufman Sample Participants

MDS with the KABC-II and KTEA-II together were based on correlations derived from participant scores who were administered the tests as part of co-norming the tests, with a testing interval of 0 to 104 days, or an average of 8 days (Kaufman and Kaufman 2004). MDS was conducted by different grade levels, and the correlations matrices for grades 1–3 (n = 592), grades 4–6 (n = 558), grades 7–9 (n = 566), and grades 10–12 (n = 401) formed four subsamples. These four subsamples were chosen because of possible developmental shifts with these data in other research (Hajovsky et al. 2014). Demographics are shown in Table 2. Subsample demographics were mostly similar to those of the entire norming sample.

Table 2. Demographic Information: Kaufman (KABC-II and KTEA-II) Subsamples, Full Sample.

Kaufman Test Demographic Information: KABC-II and KTEA-II Grade Subsamples				
	Grades 1–3	**Grades 4–6**	**Grades 7–9**	**Grades 10–12**
Sex	(n = 592)	(n = 558)	(n = 566)	(n = 401)
Female	49.3	48.9	49.5	50.9
Male	50.7	51.1	50.5	49.1
Ethnicity				
Black	15.5	13.8	15.5	13.7
Hispanic	19.9	18.3	15.4	17.2
Other	4.7	6.1	5.7	5.5
White	59.8	61.8	63.4	63.6
Highest Parent Ed.				
Grade 11 or less	13.0	16.5	14.8	15.5
HS graduate	32.6	31.9	32.2	33.4
1–3 years college	31.9	28.7	29.3	28.4
4 year degree+	22.5	22.9	23.7	22.7
Geographic Region				
Northeast	16.6	16.5	11.3	9.5
North central	23.6	27.1	23.0	27.9
South	35.5	33.2	35.0	35.9
West	24.3	23.3	30.7	26.7

Table 2. *Cont.*

Kaufman Test Demographic Information: KABC-II and KTEA-II Grade Subsamples				
	Grades 1–3	**Grades 4–6**	**Grades 7–9**	**Grades 10–12**
Age Band				
6:00–6:11	20.6			
7:00–7:11	30.1			
8:00–8:11	31.9	0.2		
9:00–9:11	16.7	16.8		
10:00–10:11	0.7	33.7		
11:00–11:11		33.3	0.2	
12:00–12:11		14.7	20.5	
13:00–13:11		1.1	32.3	
14:00–14:11		0.2	32.0	0.5
15:00–15:11			13.4	15.5
16:00–16:11			0.9	32.9
17:00–17:11			0.4	33.4
18:00–18:11			0.2	17.5
19:00–19:11			0.2	0.2

2.2. Measures

2.2.1. WISC-V and WIAT-III

The Wechsler Intelligence Scale for Children-Fifth Edition (WISC-V; Wechsler 2014) is an individually administered assessment of intelligence for children and adolescents between 6 years and 16 years, 11 months. The WISC-V provides ten primary subtest scores, six secondary subtest scores, and five complementary subtests scores. Seven of the primary subtest scores combine to form the Full-Scale IQ (FSIQ), an index of general intelligence. Pairs of the primary subtest scores combine to form five primary indexes (Verbal Comprehension, Visual-Spatial, Fluid Reasoning, Working Memory, and Processing Speed). Twenty-one of 21 subtests were included in the study. The Wechsler Individual Achievement Test, Third Edition (WIAT-III; Breaux 2009) is an individually administered measure of academic achievement for children and adolescents between the ages of 4 and 51 years, (Breaux 2009). Alphabet Writing Fluency and Early Reading Skills were excluded from analysis because they were only administered to examinees in Grade 3 or below (N = 44) so 14 of 16 subtests were included here.

2.2.2. KABC-II and KTEA-II

The Kaufman Assessment Battery for Children, Second Edition (KABC-II; Kaufman and Kaufman 2004) is an individually administered assessment of children and adolescents' processing and cognitive abilities between 3 years and 18 years, 11 months. The KABC-II provides a composite as an estimate of general intelligence and multiple CHC indexes. There were 15 of 18 subtests included in the study. The Kaufman Test of Educational Achievement, Second Edition (KTEA-II; Kaufman 2004) is an individually administered measure of academic achievement for children and adolescents between the ages of 4 years, 6 months and 25 years, 11 months (Kaufman 2004). There were 14 or 16 of 16 subtests included in the study depending on the age range.

2.3. Data Preparation Prior to MDS Analysis

MDS was conducted separately for each dataset: one with WISC-V and WIAT-III, and one with each of the four KABC-II and KTEA-IIs. Prior to MDS, symmetrical correlation matrices were created from bivariate correlations between subtest (simply "test" from now on) standard scores. From the WISC-V and WIAT-III data, 1.5% of the scores were missing, mostly from Math Fluency Multiplication and Essay Composition. From the Kaufman Grades 1–3 data (n = 592), 5.3% was missing mostly from Decoding Fluency and Word Recognition Fluency. From the Kaufman Grades 4–6 data (n = 558), Kaufman Grades 7–9 data (n = 566), and Kaufman Grades 10–12 data (n = 401), less than 1% was missing from

each. Pairwise deletion was used for calculating correlations and not considered a problem based on such small amounts of missing data (Graham 2009).

Dissimilarity matrices were constructed from the correlation matrices. The formula $\sqrt{1-r}$ was used to convert Pearson correlations for each pair of tests to dissimilarities. Each dissimilarity matrix was submitted to MDS procedures.

2.4. MDS Analysis

Five symmetrical dissimilarity matrices were inputted separately to the MDS algorithm. Initial model specifications included initial configuration (the starting location for each object in the matrix from which the algorithm produces iterations of configurations), type of transformation (whether dissimilarities are treated as interval- or rank-level data), and number of dimensions to be represented in the configuration. Configuration of objects were plotted for visual analysis and interpretation. Statistics and visualization were conducted with R 4.0.4 (R Core Team 2021), smacof (Mair et al. 2021), ggplot2 (Wickham 2016), and rgl (Adler and Murdoch 2021) packages. The output from the MDS smacof package (Mair et al. 2021) included the configuration of objects (coordinates for each point in the specified number of dimensions) and an estimate of model misfit between the dissimilarity matrix and MDS configuration, called stress. Different specifications for the model affect fit and were compared to select the best MDS model for each matrix.

2.4.1. Model Selection

Model selection was based on fit and interpretability. Four different models were estimated for each matrix: (1) interval transformation in two dimensions; (2) interval transformation in three dimensions; (3) ordinal transformation in two dimensions; and (4) ordinal transformation in three dimensions. Stress is a loss function that helps with deciding or confirming model specifics from the MDS procedure. Perfect fit results in 0 stress. The maximum stress is 1 (Cohen et al. 2006). Global stress was considered for absolute fit. Kruskal's (1964) guidelines were used as a starting point: .20 is poor, .10 is fair, .05 is good, and .025 is excellent. Stress also increases when the number of objects in MDS analysis increases, so those are not considered cutoff values. The rules for acceptable stress may be too strict for MDS with the large number of objects in each matrix in this study (Mair et al. 2016). Thus, stress from random permutations of dissimilarities were also compared to model stress. The null hypothesis is that stress in the MDS configuration from the study data is as high as or higher than stress from MDS of random dissimilarity matrices. If the null hypothesis is rejected, stress in the model is likely lower than stress in the random dissimilarity matrices—indicating acceptable stress from an absolute standpoint.

Global stress was also used to compare the relative fit of two- or three-dimensional models. The contributions of individual points to global stress (stress-per-point) were considered. In some cases a few points may account for most of the stress (Borg et al. 2018).

2.4.2. Preparation for Interpretation

Once each final model was selected, the MDS configurations were plotted using ggplot2 (Wickham 2016), and rgl (Adler and Murdoch 2021) packages in R. Test abbreviations and g-loadings were used to label tests in the MDS configuration scatterplots. In one version of each configuration scatterplot, tests were color-coded by complexity and a sphere was added to indicate the center point. In another version of the configuration scatterplot, tests were color-coded by CHC ability factor or academic achievement domain, and lines connected tests from the same CHC ability factor or academic achievement domain to ease interpretation. Last, WISC-V and WIAT-III tests were color-coded by content or by response mode to help with visual analysis and explore similarities in content and response modes.

To support visual inspections in answering questions about complexity, the center point of each MDS configuration was calculated and depicted as a black sphere. The center point was defined as the mean of each dimension $(\overline{x}, \overline{y}, \overline{z})$. The distance from each subtest

to the center point was calculated. Where the distance between $P_1 = (x_1, y_1, z_1)$ and $P_2 = (x_2, y_2, z_2)$ is calculated by $d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$.

Last, Spearman's rank-order correlation was calculated for intelligence test *g*-loadings and distance from the center point of the MDS configuration to help quantify the organization of intelligence tests in the MDS map in terms of complexity.

3. Results

3.1. Preliminary Analysis and Model Selection

Global stress for the MDS configurations across the five datasets ranged from .06 (good) to .29 (poor) depending on the matrix, type of transformation, and number of dimensions. Table 3 shows stress values for ordinal and interval transformations in two and three dimensions. Stress values in all MDS configurations were lower than the random permutations. MDS models fit the data in the dissimilarity matrices better than a configuration of random dissimilarities and were minimally acceptable in terms of absolute fit. Stress values of different model configurations were compared for relative fit. In terms of relative fit, ordinal, three-dimensional configurations fit the data best. Ordinal fit better in general because rank order of distances is preserved instead of the relative size of distances between objects, and rank order is simpler to remain consistent between the input and MDS configuration.

Last, stress-per-point values of three-dimensional MDS models were examined to identify tests that accounted for much more stress than other points. All tests were retained in the MDS configurations because the three-dimensional stress values were fair or good as is.

Table 3. Ordinal and Interval MDS Stress Comparisons in Two and Three Dimensions.

Correlation Matrix	Ordinal, Two Dimensions	Interval, Two Dimensions	Ordinal, Three Dimensions	Interval, Three Dimensions
WISC-V and WIAT-III	0.22	0.26	**0.15**	0.18
Kaufman Grades 1–3	0.24	0.29	**0.14**	0.20
Kaufman Grades 4–6	0.24	0.29	**0.14**	0.20
Kaufman Grades 7–9	0.18	0.20	**0.11**	0.18
Kaufman Grades 10–12	0.18	0.20	**0.13**	0.18

Note. MDS models with "Torgerson" classical scaling starting configuration. Bolded numbers were the lowest stress values among the four configuration for that matrix.

3.2. Primary Analyses

3.2.1. WISC-V and WIAT-III Model Results

The ordinal, three-dimensional model was selected for the WISC-V and WIAT-III data. The three-dimensional MDS configuration was plotted. The following research questions were answered based on visual analysis of the three-dimensional scatterplot (with different color-coding versions) and calculations of distances from the configuration center.

1. Are complex tests in the center of the MDS configuration with less complex tests farther from the center of the MDS configuration?

Yes and no. Intelligence and academic achievement tests, color-coded by complexity, are shown in Figure 1 (Figure 1 is also available as an interactive 3D graphic in the Supplemental Figures). Intelligence test *g*-loadings are in the test labels, when available; intelligence tests with unknown *g*-loadings are gray. The black sphere in Figure 1 is the center of the MDS configuration. Complex tests are not all in the center of the MDS map (i.e., green tests in Figure 1 are not all in the center of the map). When rotating the interactive 3D configuration, Arithmetic (highest *g*-loading of .73) is closest to the center of the MDS configuration followed by Math Problem Solving, but other complex tests (e.g., Essay Composition) are as far away as basic skills or fluency tests. Other tests with *g*-loadings

higher than .70 (Vocabulary *g*-loading of .73 and Information *g*-loading of .72) are on the periphery with tests of lower *g*-loadings. By visual inspection alone, tests do not radiate outward with the highest complexity in the center and the lowest complexity tests on the periphery.

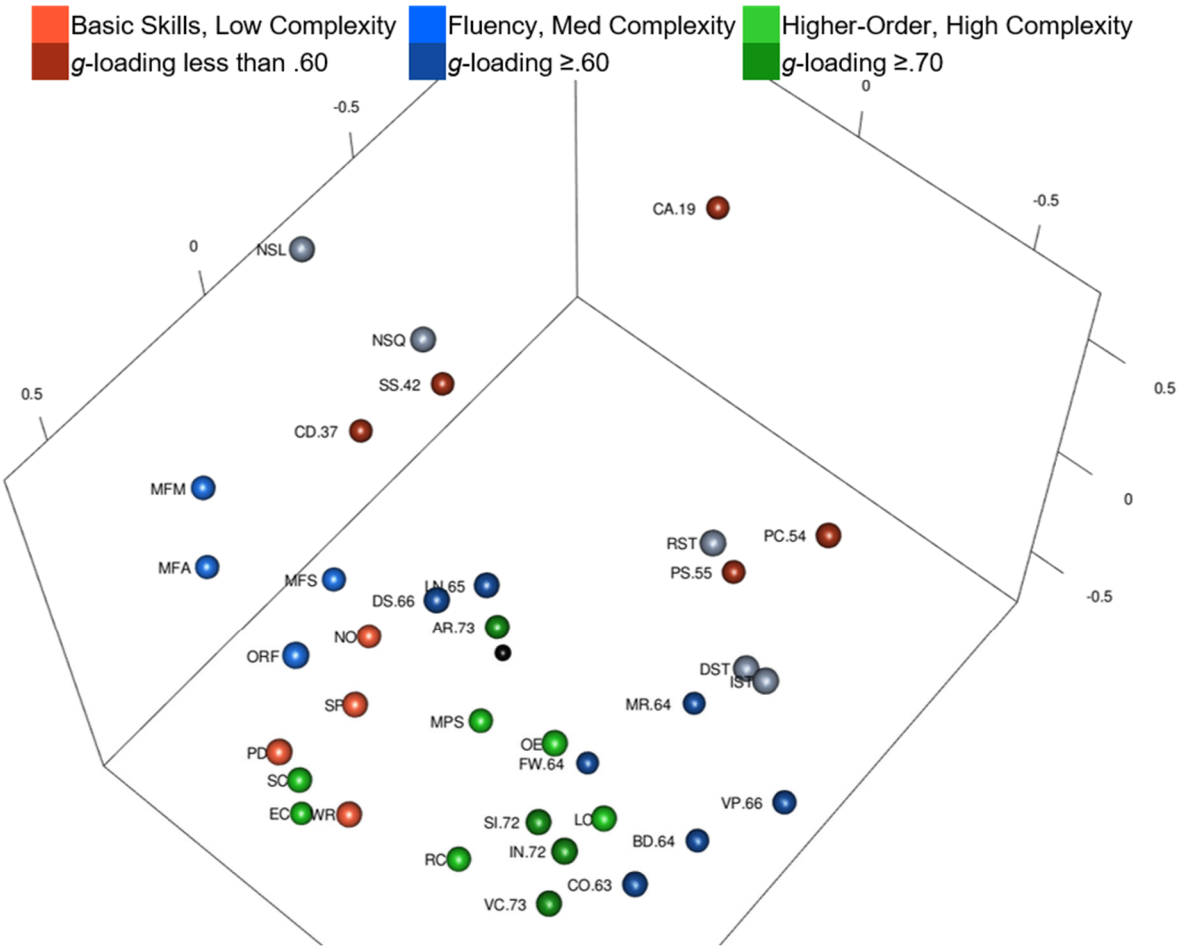

Figure 1. WISC-V and WIAT-III 3D MDS Configuration, Color-Coded by Complexity.

Tests in Table 4 are in order by distance from the WISC-V and WIAT-III MDS configuration center point. In the ranked list, Arithmetic is closest to the configuration center (.09 units from center) followed by Math Problem Solving (.24 units from center). Cancellation is the farthest from the configuration center (1.33 units from center). Intelligence and academic achievement tests are dispersed throughout the rankings (i.e., the intelligence tests are not all at the top or bottom of the rank order). Several high complexity intelligence and academic achievement tests are within .6 units from the center, but Essay Composition, Matrix Reasoning, Block Design, and other higher complexity tests are as far from the center as lower complexity intelligence and academic achievement tests. A Spearman's rank-order correlation was calculated to assess the relation between intelligence test complexity (*g*-loadings when available) and distance from the MDS configuration center. There was a statistically significant, strong negative correlation between intelligence test *g*-loadings and distance from the configuration center, $rs(14) = -.715$, $p < .01$. Intelligence tests with higher *g*-loadings were more likely to be closer to the center of the configuration, but those are only based on intelligence tests with *g*-loadings.

Table 4. WIAT-III and WISC-V Subtests Ordered by Distance from Center of 3D Configuration.

Subtest Abbr.	Subtest	Composite	*g*-Loading or Complexity	Distance from Center
AR	Arithmetic	Fluid Reasoning	.73	0.09
MPS	Math Problem Solving	Mathematics	High	0.24
LN	Letter-Number Sequencing	Working Memory	.65	0.26
NO	Numerical Operations	Mathematics	Low	0.39
SP	Spelling	Written Expression	Low	0.40
DS	Digit Span	Working Memory	.66	0.40
OE	Oral Expression	Oral Language	High	0.43
SI	Similarities	Verbal Comprehension	.72	0.45
LC	Listening Comprehension	Oral Language	High	0.50
RC	Reading Comprehension	Reading Comp. & Fluency	High	0.50
MFS	Math Fluency Subtraction	Math Fluency	Medium	0.53
WR	Word Reading	Basic Reading	Low	0.56
SC	Sentence Composition	Written Expression	High	0.56
VC	Vocabulary	Verbal Comprehension	.73	0.62
IN	Information	Verbal Comprehension	.72	0.63
CO	Comprehension	Verbal Comprehension	.63	0.63
DST	Delayed Symbol Translation	Symbol Translation		0.66
PD	Pseudoword Decoding	Basic Reading	Low	0.67
ORF	Oral Reading Fluency	Reading Comp. & Fluency	Medium	0.68
RST	Recognition Symbol Translation	Symbol Translation		0.68
PS	Picture Span	Working Memory	.55	0.70
IST	Immediate Symbol Translation	Symbol Translation		0.74
FW	Figure Weights	Fluid Reasoning	.64	0.78
NSQ	Naming Speed Quantity	Naming Speed		0.78
BD	Block Design	Visual Spatial	.64	0.80
MFA	Math Fluency Addition	Math Fluency	Medium	0.80
CD	Coding	Processing Speed	.37	0.81
VP	Visual Puzzles	Visual Spatial	.66	0.82
MR	Matrix Reasoning	Fluid Reasoning	.64	0.82
PC	Picture Concepts	Fluid Reasoning	.54	0.82
SS	Symbol Search	Processing Speed	.42	0.82
MFM	Math Fluency Multiplication	Math Fluency	Medium	0.85
EC	Essay Composition	Written Expression	High	0.86
NSL	Naming Speed Literacy	Naming Speed		1.06
CA	Cancellation	Processing Speed	.19	1.33

2. Are intelligence tests and academic achievement tests clustered by CHC ability and academic content, respectively?

Yes. WISC-V and WIAT-III MDS CHC ability and academic domain clusters are visible in Figure 2 (Figure 2 is also available as an interactive 3D graphic in the Supplemental Figures). Tests are color-coded by CHC ability or academic achievement domain and lines connect the clusters of tests. Clusters are generally separate (e.g., there are no math tests inside of the reading cluster) and clusters of tests are mostly separated by their different color-coding. Reading, writing, math tests in the WISC-V and WIAT-III 3D MDS configuration are in a separate wedge from intelligence tests (on the right-hand side in

Figure 2). The writing cluster is between the reading cluster and the math cluster. The oral language cluster is separate from the other academic achievement areas, and much closer to the Gc cluster as shown in Figure 2. Math Problem Solving is closer than reading tests to the Gc cluster. Math tests are closer than reading and writing tests to the Gf cluster. The Math Problem Solving test involves similar questions and skills as the Arithmetic (Gf) test and requires higher-order thinking and problem solving like Arithmetic and the other fluid reasoning tests.

Even though the CHC and academic achievement clusters do not overlap (e.g., there is not a Gwm test within the cluster of Gc tests), some of the tests are near other clusters. Arithmetic is part of the Gf cluster, but it is located near the math cluster. Picture Span is part of the Gwm cluster, but it is located near the Picture Concepts test, a test with similar picture content (instead of letter and numeric content like the other Gwm tests). Oral Reading Fluency is in the reading cluster, but it is located near Digit Span and Letter-Number Sequencing Gwm tests. Gs and NSI clusters (lower left in Figure 2) are isolated from other clusters. Those tests have the lowest correlations with other tests. Lastly, Gf and Gv tests associated with primary WISC-V indexes were in separate clusters in the WISC-V and WIAT-III map.

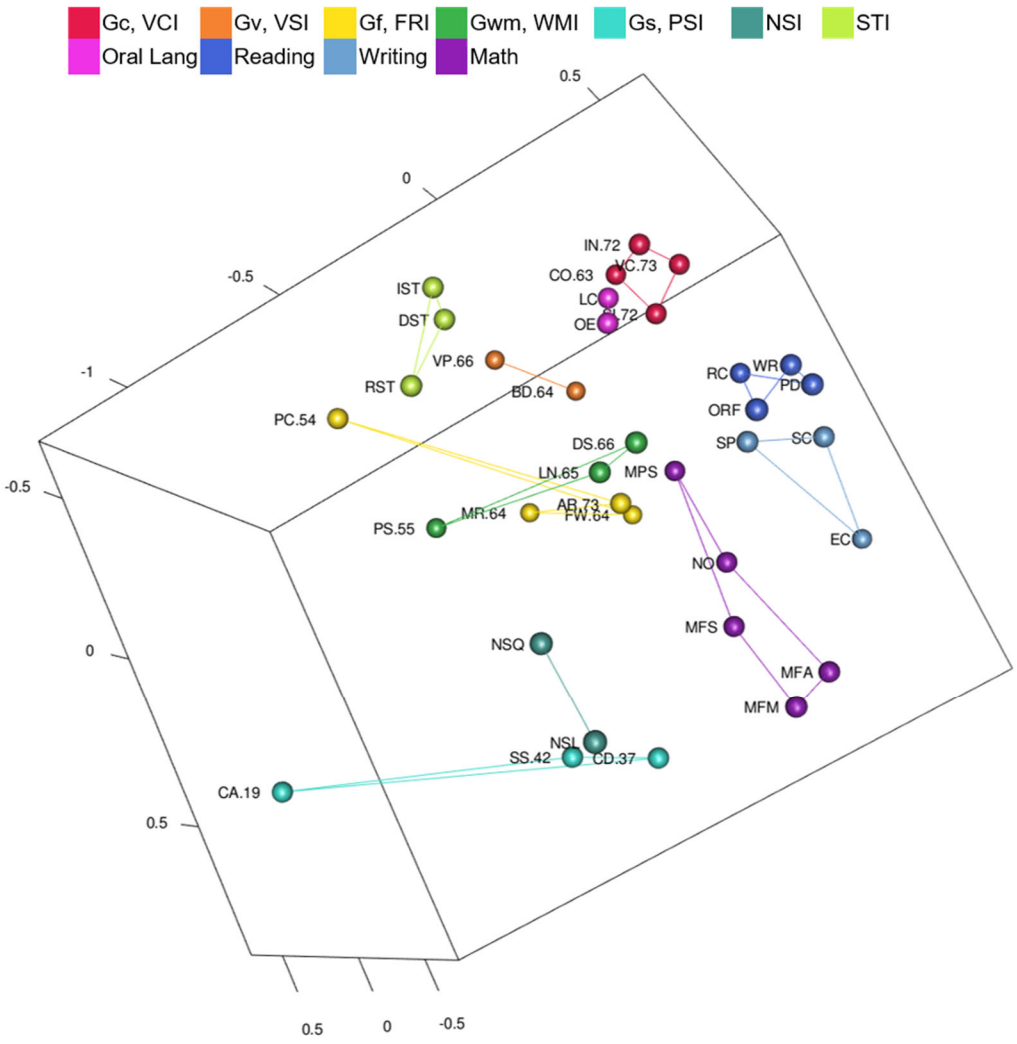

Figure 2. WISC-V and WIAT-III 3D MDS Configuration CHC and Academic Clusters.

3. Are tests organized into auditory-linguistic, figural-visual, reading-writing, quantitative-numeric, and speed-fluency regions?

Yes, though some regions overlap. Tests are color-coded by CHC ability or academic achievement domain, and dashed ovals highlight the auditory-linguistic, figural-visual, reading-writing, quantitative-numeric, and speed-fluency regions in Figure 3. The auditory-linguistic region includes Gc and oral language tests. The reading-writing region includes reading and writing tests. The quantitative-numeric region includes math achievement tests, memory tests with number stimuli, and fluid reasoning tests in which examinees *solve* novel math problems. The figural-visual region includes tests with pictures or other visual stimuli. The speed-fluency region includes processing speed, rapid naming, and math fluency tests. Math fluency tests are in the overlap between quantitative-numeric and speed-fluency regions.

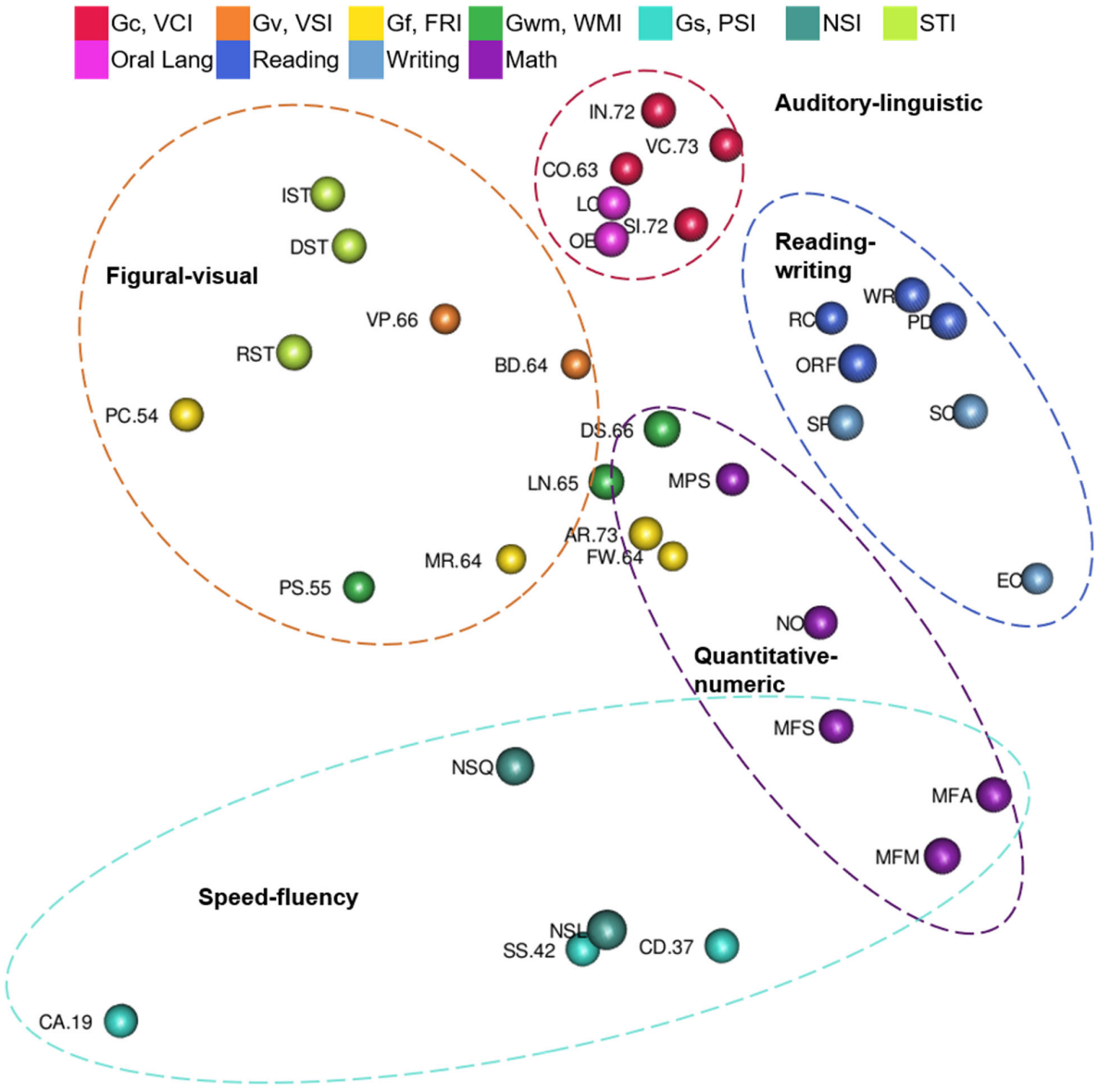

Figure 3. WISC-V and WIAT-III 3D MDS Configuration Regions.

3.2.2. Kaufman Grades 4–6 Model Results

Only one Kaufman model was included due to space limitations. Similarities and differences across grade bands, however, will be discussed. See the Supplementary File for three-dimensional figures of the other Kaufman models. The ordinal, three-dimensional model was selected for the Kaufman Grades 4–6 data. The three-dimensional MDS configuration was plotted. The following research questions were answered based on visual

analysis of the three-dimensional scatterplot (with different color-coding versions) and calculations of distances from the configuration center.

1. Are complex tests in the center of the MDS configuration with less complex tests farther from the center of the MDS configuration?

Yes and no. Intelligence and academic achievement tests are color-coded by complexity in Figure 4 (Figure 4 is also available as an interactive 3D graphic in the Supplemental Figures). Intelligence tests' *g*-loadings are in the test labels, when available. The black sphere in Figure 4 is the center of the MDS configuration. Three complex academic achievement tests are closest to the black sphere: Reading Comprehension, Written Expression, and Math Concepts & Applications. Letter & Word Recognition (a low complexity test) is also close to the center.

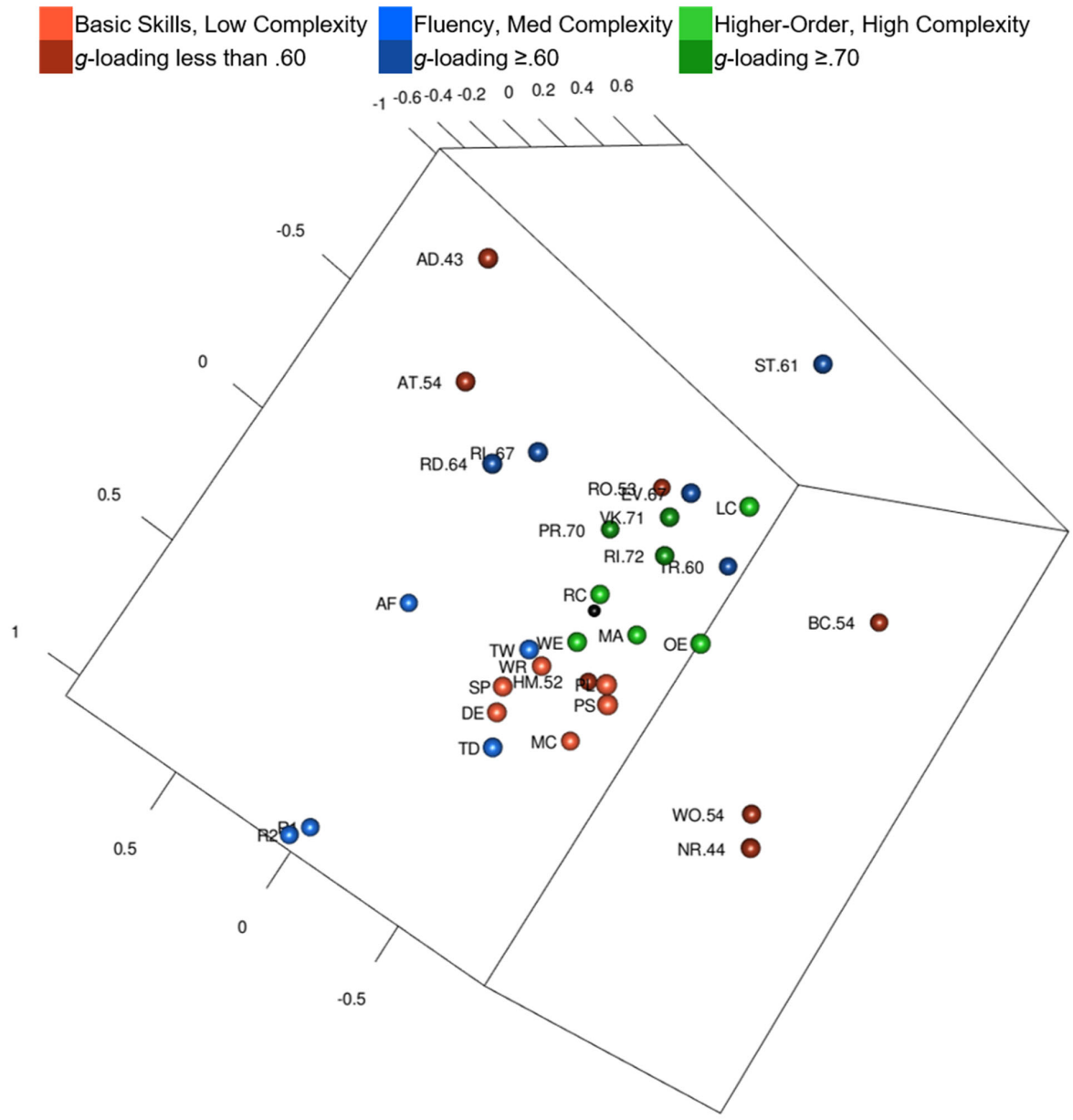

Figure 4. Kaufman Grades 4–6 3D MDS Configuration, Color-Coded by Complexity. Kaufman Grades 4–6 3D MDS Configuration Static View, Color-Coded by Complexity.

Tests in Table 5 are in the order by distance from the Kaufman Grades 4–6 MDS configuration's center point. Reading Comprehension (.05 units from center), Math Concepts

& Applications (.18 units from center), and Written Expression (.21 units from center) are closest to center. Riddles (.72) and Verbal Knowledge (.71) are the tests with the highest *g*-loadings; they are the intelligence tests closest to center. There was a statistically significant, strong negative correlation between intelligence test *g*-loadings and distance from center, $rs(13) = -.842$, $p < .01$. Intelligence tests with higher *g*-loadings were more likely to be closer to center than intelligence tests with lower ones. Tests with *g*-loadings $\geq .70$ are closest to the center of the configuration, followed by tests with *g*-loadings > .61, and then tests with *g*-loadings $\leq .61$. Intelligence tests are organized with more complex tests closer to center and less complex tests around the periphery, but again, seven out the ten tests closest to the center are achievement tests and not intelligence tests.

Table 5. Kaufman Subtests Ordered by Distance from Center of 3D Configuration (Grades 4–6).

Subtest Abbr.	Subtest	Composite	*g*-Loading or Complexity	Distance from Center
RC	Reading Comprehension	Reading	High	0.05
MA	Math Concepts & Applications	Mathematics	High	0.18
WE	Written Expression	Written Language	High	0.21
WR	Letter & Word Recognition	Reading, Decoding	Low	0.24
TW	Word Recognition Fluency	Reading Fluency	Medium	0.27
RI	Riddles	Gc	0.72	0.28
VK	Verbal Knowledge	Gc	0.71	0.37
SP	Spelling	Written Language	Low	0.41
PR	Pattern Reasoning	Gf	0.7	0.43
DE	Nonsense Word Decoding	Sound-Symbol, Decoding	Low	0.45
MC	Math Computation	Mathematics	Low	0.46
EV	Expressive Vocabulary	Gc	0.67	0.47
OE	Oral Expression	Oral Language	High	0.47
TD	Decoding Fluency	Reading Fluency	Medium	0.53
RL	Rebus	Glr	0.67	0.56
LC	Listening Comprehension	Oral Language	High	0.61
RD	Rebus Delayed	Glr	0.64	0.63
TR	Triangles	Gv	0.6	0.64
PS	Phonological Awareness	Sound-Symbol	Low	0.66
PL	Phonological Awareness (Long)	Sound-Symbol	Low	0.69
AF	Associational Fluency	Oral Fluency	Low	0.77
HM	Hand Movements	Gsm	0.52	0.79
AT	Atlantis	Glr	0.54	0.8
WO	Word Order	Gsm	0.54	0.8
NR	Number Recall	Gsm	0.44	0.86
RO	Rover	Gv	0.53	0.89
ST	Story Completion	Gf	0.61	1.03
BC	Block Counting	Gv	0.54	1.06
AD	Atlantis Delayed	Glr	0.43	1.12
R2	Naming Facility: Objects, Colors, & Letters	Oral Fluency	Low	1.21
R1	Naming Facility: Objects & Colors	Oral Fluency	Low	1.22

2. Are intelligence tests and academic achievement tests clustered by CHC ability and academic content, respectively?

Yes. Kaufman Grades 4–6 CHC ability and academic domain clusters are visible in in Figure 5 (Figure 5 is also available as an interactive 3D graphic in the Supplemental Figures). Tests are color-coded by CHC ability or academic achievement domain, and lines connect the clusters of tests. The writing cluster is inside of the reading cluster. The reading-writing cluster is near math academic achievement tests. Oral language tests are split into two clusters on either side of academic achievement test clusters. Oral Expression and Listening Comprehension are next to the Gc cluster. Naming Facility and Associational Fluency are on the other side of the MDS map. Math Concepts & Applications is closer to Gf tests, but Math Computation is not closer to Gf tests than are the reading tests. Most reading tests, except Reading Comprehension, are farther from Gc than Math Concepts & Applications.

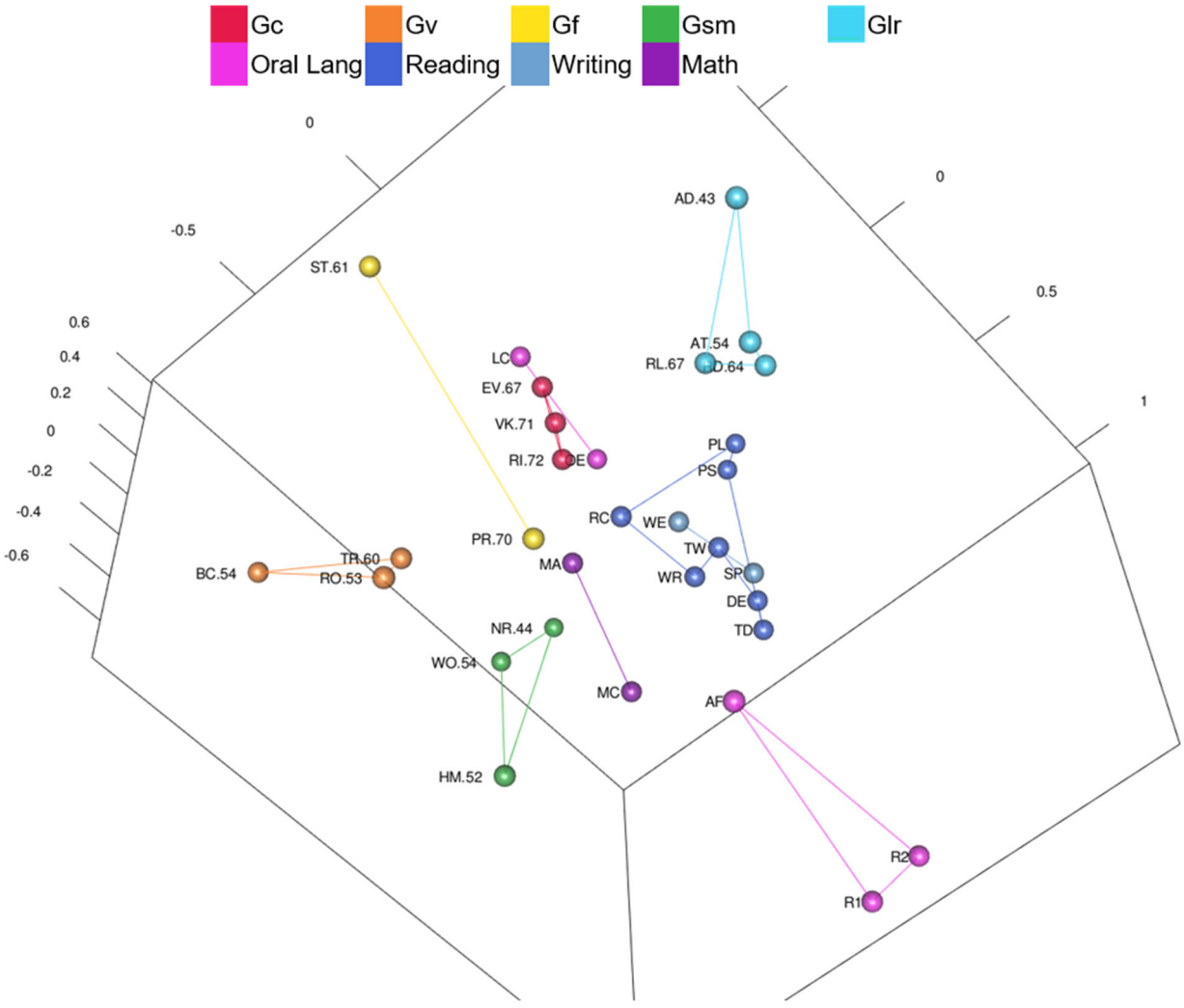

Figure 5. Kaufman Grades 4–6 3D MDS Configuration CHC and Academic Clusters.

3. Are tests organized into auditory-linguistic, figural-visual, reading-writing, quantitative-numeric, and speed-fluency regions?

Yes, but not speed-fluency. Tests are color-coded by CHC ability or academic achievement domain, and dashed ovals highlight the auditory-linguistic, figural-visual, reading-writing, and quantitative-numeric regions in Figure 6. Naming Facility tests and Associational Fluency were not in the auditory-linguistic region with Oral Expression and

Listening Comprehension tests. Thus, those tests formed a small retrieval fluency region. The reading-writing region includes reading and writing tests. The two math academic achievement tests are near the reading-writing region and form the quantitative-numeric region. The figural-visual region includes tests with pictures or other visual stimuli. Tests in the figural-visual region are spread out more than the tests in other regions of the configuration.

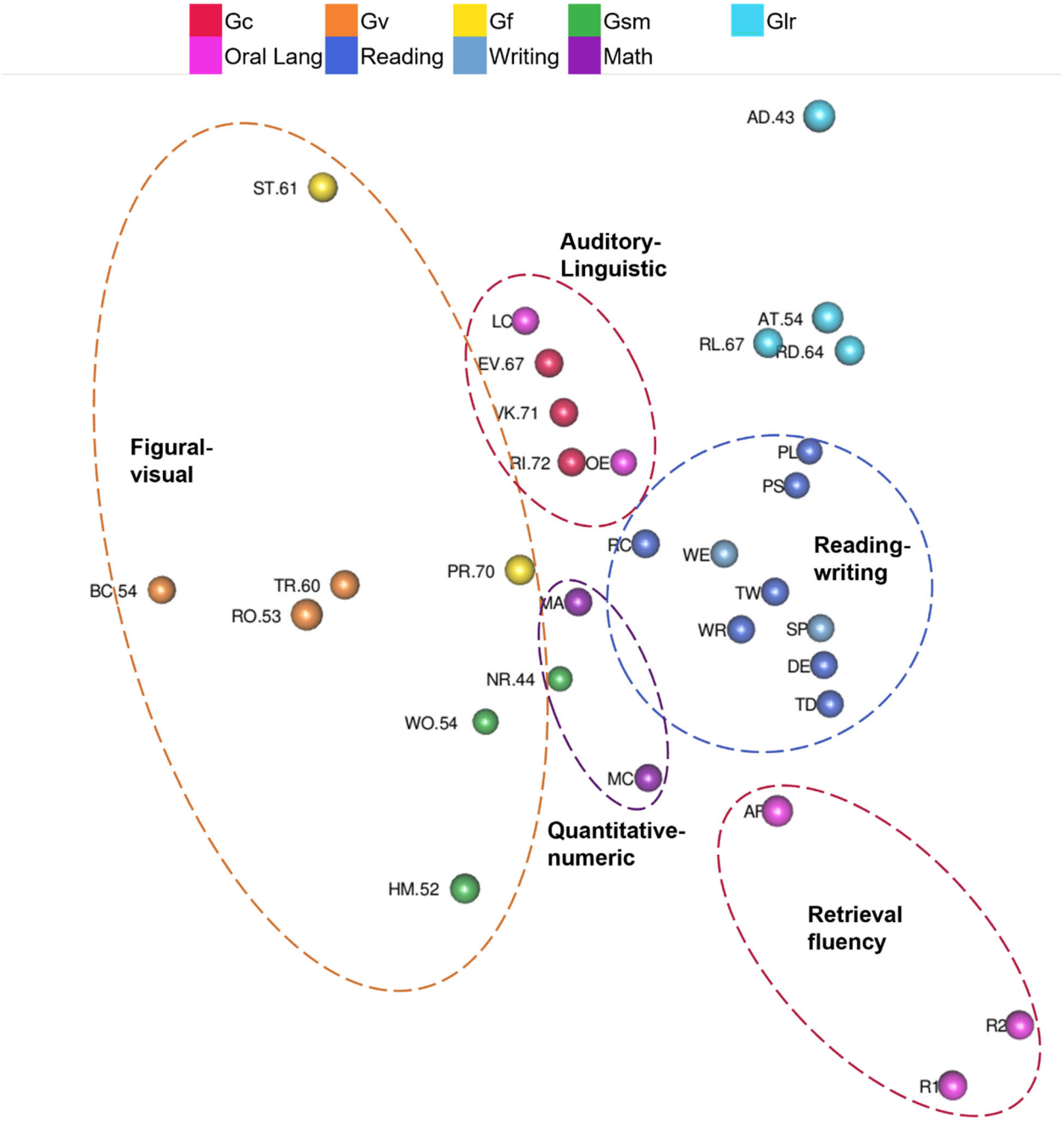

Figure 6. Kaufman Grades 4–6 3D MDS Configuration Regions.

3.3. Secondary Analyses

3.3.1. Kaufman Grade Groups

MDS results were mostly similar across grade groups, but there were some visible differences. Grade groups were similar in that among intelligence tests, those with higher *g*-loadings were closer to the center of the MDS configuration. They were also similar in that the tests closest to the center were more likely to be achievement tests than intelligence tests. The pattern of lower complexity tests radiating outward toward the periphery of the configuration was not as obviously visible. For example, Letter & Word Recognition was second closest to the center of the Grades 10–12 MDS configuration.

Another consistent finding was that Gc, reading, writing, and math clusters tended to be in the center of MDS configurations for all grade groups. One potential developmentally related finding, however, was that Listening Comprehension and Oral Expression were closer to the center with the previously mentioned clusters at the upper grade levels (Grades 7–9 and Grades 10–12).

Additionally, the 3D MDS configuration regions were similar across grade groups. The maps all included clearly delineated auditory-linguistic, figural-visual, reading-writing, and quantitative-numeric regions.

One potential developmental finding was related to where fluency tests from the oral language composites were located on the maps. In Grades 1–3 and 4–6, these fluency tests were on the opposite side of the MDS configuration from the other auditory-linguistic tests. In Grades 7–9 and 10–12, however, these fluency tests were on the same side of the MDS configuration as other auditory-linguistic tests.

One other notable difference was found, but it was only related to one grade group. In each grade group, Word Recognition Fluency and Decoding Fluency were much closer to reading tests (clustered by academic content instead of by fluency); however, both of these fluency tests were on the side of the reading cluster closest to the Associational Fluency and Naming Facility tests in the Grades 7–9 MDS configuration. There was only a memory region in the Kaufman Grades 10–12 configuration. In the other grade configurations, Glr and Gsm tests were on opposite sides of the configuration or very far from each other.

3.3.2. WISC-V and WIAT-III Content and Response Modes

Guttman and Levy (1991) interpreted two additional facets beyond complexity: content and response mode. In order to explore Guttman's content facet in the WISC-V and WIAT-III model, tests were color-coded more broadly by the type of content or stimuli they include. This coding scheme is shown in Figure 7 (Figure 7 is also available as an interactive 3D graphic in the Supplemental Figures). Tests with verbal and figural content are clearly separated from each other and located in non-overlapping regions of the MDS configuration. Tests with numeric content also appear clustered closer together and in a region of their own. Some of the tests, like Letter-Number Sequencing or Coding, have combinations of different stimuli like symbols, letters, and numbers. The mixed stimuli tests are located throughout the MDS configuration, mostly between the numeric and figural content regions. Taken together and as shown clearly in Figure 7 the content facet is useful in analyzing larger patterns in the MDS configuration.

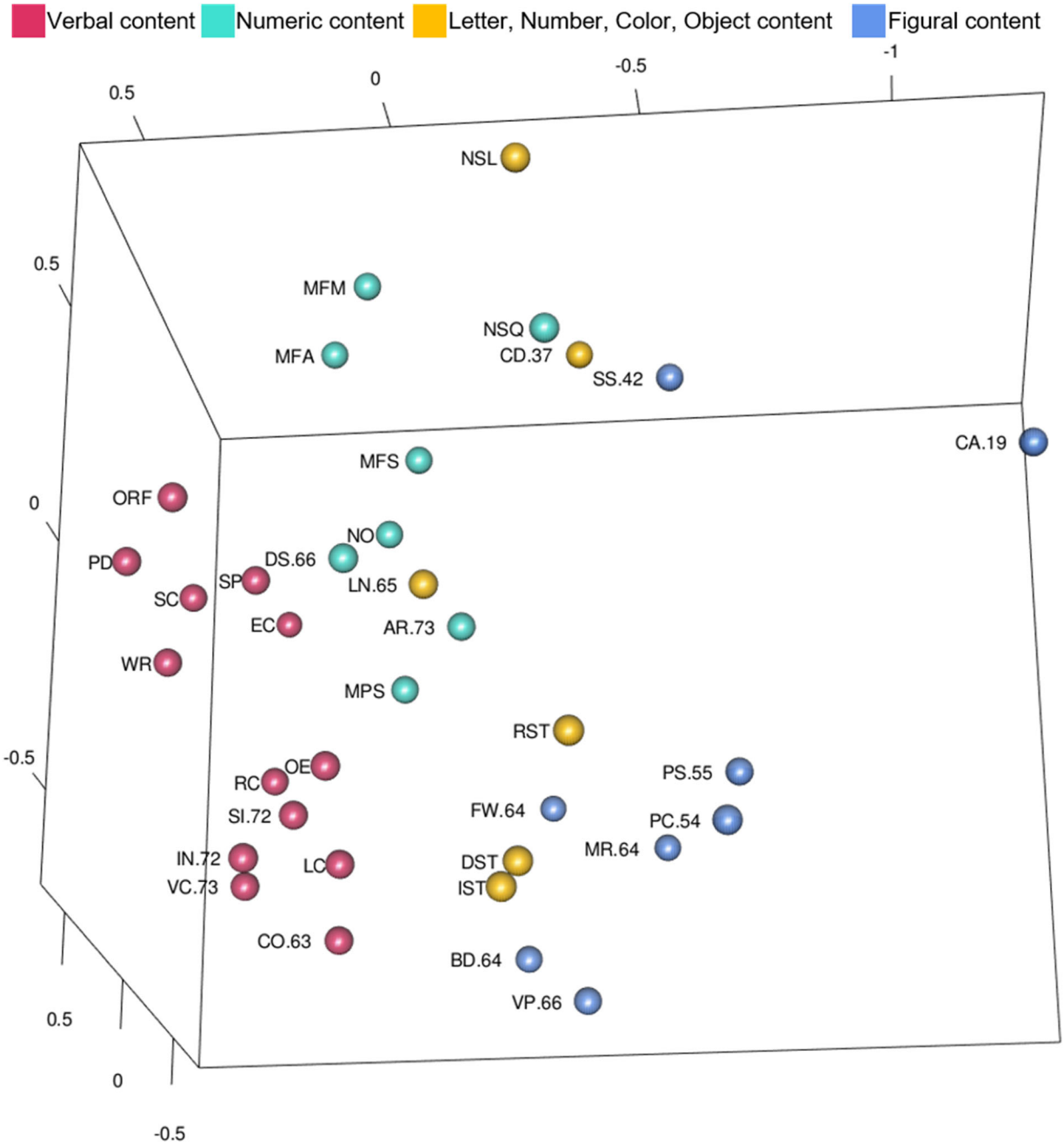

Figure 7. WISC-V and WIAT-III 3D MDS Configuration Static View, Color-Coded by Content.

In order to explore Guttman's response mode facet in the WISC-V and WIAT-III model, tests were color-coded by the way an examinee responds and these are shown in Figure 8 (Figure 8 is also available as an interactive 3D graphic in the Supplemental Figures). Notably, tests with paper-pencil responses are mostly on one side of the MDS configuration and tests with verbal responses are mostly on the other side. Tests that examinees may respond verbally to or by pointing to are on the periphery of the MDS configuration (the lowest portion of Figure 8). Block Design is the only test in which a manual response (moving blocks) is required, and it is located near the tests with manual or verbal response options.

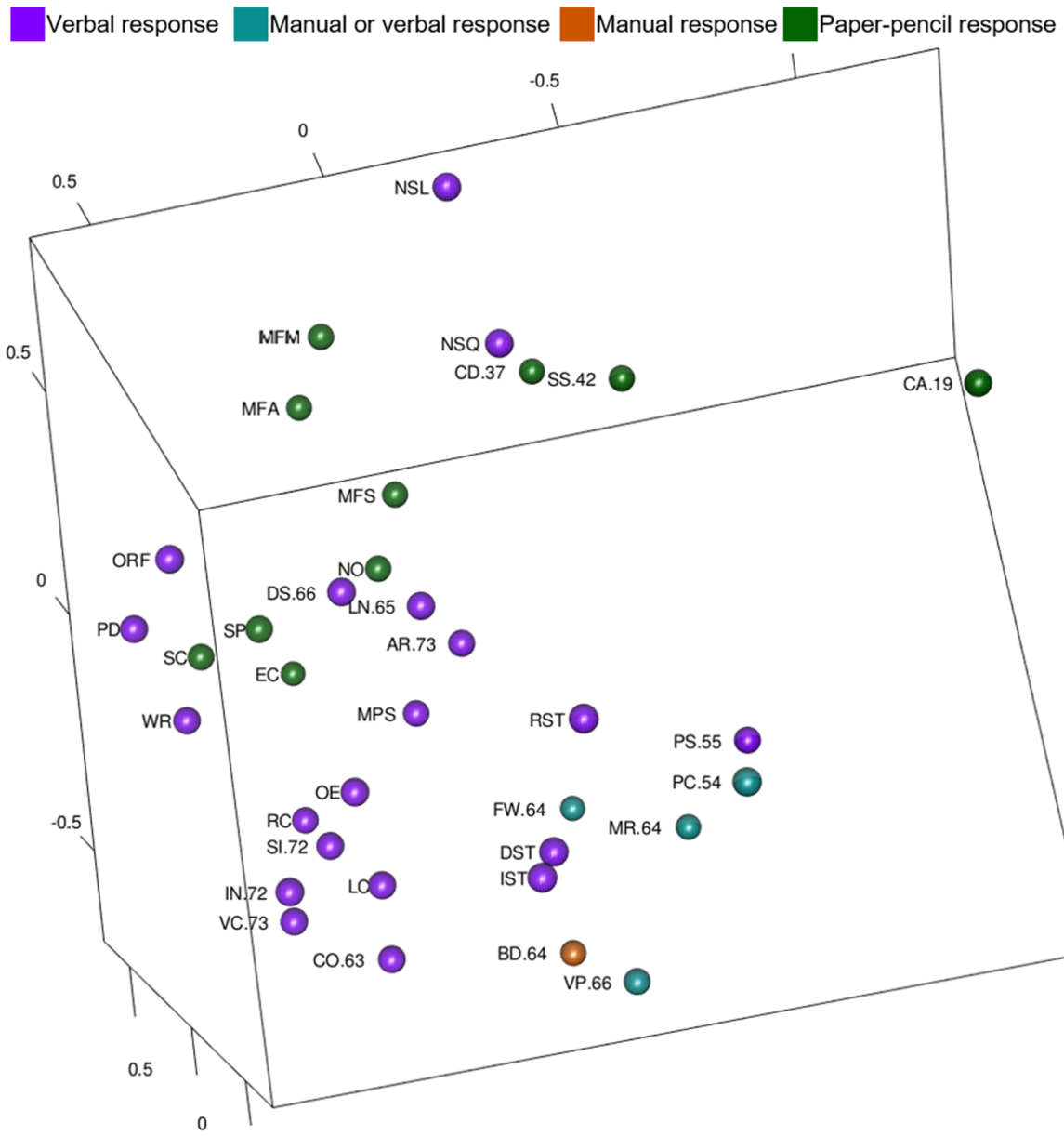

Figure 8. WISC-V and WIAT-III 3D MDS Configuration, Color-Coded by Response Mode.

4. Discussion

The purpose of this study was to use MDS to analyze correlations among Wechsler cognitive and achievement tests and among Kaufman cognitive and achievement tests to better understand the relations among the scores. Three research questions were answered.

First, less complex academic achievement tests were in the center of the MDS configurations with complex academic achievement tests and intelligence tests with high *g*-loadings. Intelligence tests with high *g*-loadings were more likely to be near the center than intelligence tests with lower *g*-loadings. However, academic achievement tests were more likely to be near the center of the configuration than intelligence tests. Fluency tests were least likely to be near the center of the configuration. The finding made the complexity interpretation less clear and was an unexpected result.

Second, intelligence and academic achievement tests were generally clustered by CHC ability and academic content, respectively. Reading, writing, complex oral language, and Gc tests were consistently clustered near each other. Complex math tests were closer to Gf tests than simpler math tests, but the math cluster was not always closer to the Gf tests

than to Gc tests. CHC abilities were helpful in explaining the locations of intelligence constructs and academic achievement content areas were useful in explaining the academic area clusters.

Third, consistent with McGrew et al. (2014), tests were organized into auditory-linguistic, figural-visual, reading-writing, quantitative-numeric regions for all model results. Speed-fluency was more visible in WISC-V and WIAT-III results than the Kaufman results. However, Kaufman tests do not include processing speed tests, and in each Kaufman model there was a retrieval fluency region. At times, the regions were overlapping as some fluency tests clustered with speeded tests in a speed-fluency region and others clustered with tests of similar academic content. For example, math fluency tests were in the overlap between the quantitative-numeric and speed-fluency regions of the WISC-V and WIAT-III model.

In addition to the research questions, most MDS results were similar across grade groups (complexity organization, Gc tests near the center, separation of complex oral language from oral fluency tests). There were, however, a few differences, with at least two possible developmental differences. Lastly, the Wechsler tests were organized by content and response processes.

Although several findings were expected, not all of our original hypotheses were confirmed. Regarding complexity, the addition of achievement tests appeared to cloud the interpretation of that dimension. As with prior research CHC classification works and confirmed McGrew et al.'s (2014) additional categorization of broader regions. Last, found that response processes may explain some of the correlations among tests. Additionally, some test specific findings emerged. We discuss these findings below.

4.1. Complexity

Complexity of a task refers to the number of cognitive processes involved, the importance of cognitive processes, attention and memory demands, and adaptation or executive functions required (Lohman and Lakin 2011). Here, for intelligence tests we also used *g*-loadings as indicators of complexity. For achievement tests, skill in more complex tasks tend to rely on skills measured in the less complex tasks. Previous studies with MDS of intelligence test score correlations have supported the radex model of intelligence, with complex tests near the center of two-dimensional or three-dimensions MDS configurations (Guttman and Levy 1991; Marshalek et al. 1983).

4.1.1. Wechsler Models

The WISC-V and WIAT-III model told a complicated story with no ending regarding complexity. Some findings were consistent with prior research. For example, in general, intelligence tests with higher *g*-loadings were closer to the center (Snow et al. 1984). The location of Arithmetic in the three-dimensional map with intelligence and achievement tests analyzed together was consistent with prior MDS research with the entire WISC-V standardization sample in two dimensions (Meyer and Reynolds 2018). Arithmetic, which also had the highest *g*-loading, was very near the configuration center. Arithmetic is classified as a Gf test according to the WISC-V manual (Wechsler 2014). With the complex math test (Math Problem Solving) being the next closest to the center, at first glance these findings appeared consistent with Marshalek et al.' (1983) findings that showed Gf tests in the center of their MDS map of WAIS and other intelligence tests, with reading and math composites close by. A closer look, however, revealed that in the current study, the remaining Gf tests were some of the furthest from the center. This finding is not only inconsistent with Marshalek and colleagues' findings, but also with the findings from factor analysis that *g* and Gf are statistically indistinguishable (Gustafsson 1984). That is, "*g*" in MDS is associated with the center of the MDS map. In addition to Arithmetic, other tests that were closest to the center in the current study were two math tests (Numerical Operations and Math Problem Solving), two working memory tests (Digit Span and Letter-Number Sequencing), and Spelling. Besides Arithmetic and Math Problem Solving, these

other tests are not considered the most complex. These working memory test locations in the configuration are inconsistent with findings from previous research in which memory tests (Digit Span Forward, Digit Span Backward, Auditory Letter Span, and Visual Number Span) were the furthest away from the center (Marshalek et al. 1983). Here, complexity did not seem to be the reason these tests were closest to the center with WISC-V and WIAT-III tests. Besides Spelling, all of the tests in the center of the WISC-V and WIAT-III MDS configuration involve numbers, so that may explain why they were closer to each other, but it breaks from prior research (Cohen et al. 2006; Guttman and Levy 1991; Marshalek et al. 1983; Meyer and Reynolds 2018) in that something besides complexity is explaining why tests are located in the center of the model—almost appearing to be test content related. Returning to Arithmetic, the constructs measured by Arithmetic have been debated, although Gf, Gwm, and math reasoning have all been implicated, so it is notable that it was close to other Gwm and mathematics tests in this study (Keith and Reynolds 2010).

The remaining WISC-V and WIAT-III tests also failed to show a pattern like the radex model with low complexity tests around the periphery and high complexity tests in the center. Instead, high and low complexity tests were intermixed in their distances from the center. For example, complex reading, oral language, writing, and Gc tests were farther away from the center even though these are considered some of the most complex tests. Although complexity has been described as a "modulating" facet that determines a test's distance from the center in a MDS map (Guttman and Levy 1991, p. 97), the pattern of intelligence and achievement tests in combination and in relation to their distances from the center in this study almost seemed to arranged by content features.

4.1.2. Kaufman Models

The Kaufman MDS model organization by content seemed more apparent than organization by complexity. At the surface, the Kaufman MDS map appeared to follow findings related to complexity and *g*-loadings (Marshalek et al. 1983; Snow et al. 1984; Tucker-Drob and Salthouse 2009). Intelligence tests correlated almost perfectly with distance from the center of the configuration and *g*-loadings, but upon closer inspection, it was only intelligence tests relative to other intelligence tests in terms of *g*-loadings that were closest to the center. Overall, the achievement tests were closer (both complex and not complex) to the center. This finding is counter to Marshalek et al.'s (1983), but in that study, only academic achievement composites were included so less complex academic achievement tests could not be in the center of the configuration. In the Snow et al. (1984) MDS analysis with Thurstone's (1938) ability data, none of the simple math tests (addition, subtraction, and multiplication) were in the center of the MDS map, but it is unknown whether these tests were simple and timed, like the fluency tests, or like the basic calculation tests in the present study. In the current study, basic math calculation and math problem solving tests were closer to the center of the map.

One clear feature of the Kaufman MDS maps was that reading, writing, math, and oral language comprehension tests were either near or intermediate distances from the center of the configuration. Across the grade bands of Kaufman MDS configurations, high complexity academic achievement tests (Math Concepts & Applications, Reading Comprehension, and Written Expression) were near the center. Gc tests were often near the center too. Less complex reading academic achievement tests from the KTEA-II were close to the center of the MDS configuration (e.g., Letter & Word Recognition was fourth from the center for Grades 1–3 and Grades 4–6, second from the center for Grades 7–9 and Grades 10–12). Nevertheless, exceptions to the radex organization, like Letter & Word Recognition and Numerical Operations, stand out when analyzing by visual inspection. The verbal centric Kaufman MDS configurations with Gf, Gv, Glr, and Gsm tests farther away from the Gc, reading, writing, and math tests called back to Vernon's (1950) hierarchical model with verbal:educational and spatial:mechanical group factors and Cattell's (1943) fluid intelligence and crystallized in intelligence. In this study, the verbal:educational

tests (crystallized intelligence) were in the center of the Kaufman configurations and spatial:mechanical (fluid intelligence) tests were to the side and periphery of the models.

One take home from these findings may be that intelligence tests do tend to generally emanate outward from the center in a way that aligns with tests with higher *g*-loadings being closer to the center. However, when considering the realm of all tests measuring cognitive abilities and developed achievement areas, the complexity dimension is not as clear. Thus, although findings are generally consistent with prior studies with intelligence tests only, it is likely premature to consider that dimension as one that is directly related to the complexity of the task.

4.2. CHC and Academic Clusters

CHC theory is framework of latent cognitive abilities (Keith and Reynolds 2010) and academic domains, like reading, writing, and quantitative knowledge. Though the interpretation of complexity was unclear, organization of subtests by CHC (Gc, Gv, Gf, Gwm or Gsm, Gs, if applicable, and Glr, if applicable) and academic domain (reading, writing, math, oral language) was very clear in all of the MDS plots. The findings in general appeared to be consistent with factor analytic evidence (Reynolds and Keith 2007; Reynolds and Keith 2017).

It was helpful to examine the MDS results through the lens of CHC factor structure and the scoring structure of academic achievement tests to subdivide the geometric space. These clusters are a succinct way to summarize results and are useful for quickly finding tests in a visually dense representation like the 3D MDS configurations. CHC and academic clusters were very consistently aligned with CHC factors and academic domains. Some of the clusters, like WISC-V Gf and Gs clusters or the KABC-II Gv cluster were more spread out in geometric space, but the lines connecting CHC and academic clusters never made a messy web of lines. The clusters, except for writing and reading, were distinct from each other.

Similar to McGrew et al. (2014) results, tests clustered by CHC and academic domain. Tests in the Wechsler and Kaufman CHC and academic domain clusters stayed together as predicted, with the exception of the oral language tests and oral fluency tests separating. In CHC parlance, however, oral retrieval fluency tests in the KTEA-3 (Kaufman et al. 2014) are known as tests of ideational fluency and rapid naming, the latter of which is also measured by the Naming Speed tasks in the WISC-V. Oral retrieval fluency tests from the KTEA-II were not close to the other oral language tests in the auditory-linguistic region, and seemed to form a separate retrieval fluency region. The retrieval fluency region was similar to the speed-fluency region found in the WISC-V and WIAT-III maps, except no tests of processing speed are included in the Kaufman tests. This fluency region was also similar to the retrieval fluency factor, Gr, recently identified as separate from learning efficiency (Jewsbury and Bowden 2016; Schneider and McGrew 2018).

In addition to clear definition of the geometric space in terms of CHC abilities and academic domains, there were interesting findings related to the tests themselves. For example, the WISC-V scoring structure separates tests into Fluid Reasoning and Visual Spatial indexes (Wechsler 2014), even though WISC-V fluid reasoning tests require examinees to reason with visual content. The separation of Gv and Gf clusters in the WISC-V and WIAT-III map supported separate composites for the primary indexes. At the same time, shared visual content that contributes to Gv and Gf clusters being near each other in the map is worth considering more carefully (cf., Reynolds and Keith 2017).

In addition, WISC-V Picture Span was part of the Gwm cluster (and included on the Working Memory Index on the test), but it was located near the Picture Concepts test that has similar picture content (instead of letter and numeric content like the other Gwm tests). When Picture Span and Picture Concepts were analyzed in a two-dimensional MDS configuration in previous research, they were across from, and not next to, each other in the MDS map (Meyer and Reynolds 2018). The current study, however, included more dimensions and academic achievement tests. A potential explanation for different findings

comes from Guttman's content facet (shown in Figure 7). The inclusion of academic achievement tests in the WISC-V and WIAT-III 3D MDS configuration introduced several tests with verbal content (reading and writing tests with Gc and some oral language tests on the left side of Figure 7). There were also more tests with numeric content (to the right and above verbal content tests in Figure 7). Including WIAT-III scores did not contribute additional tests with primarily symbols, pictures, or figure content. It is possible that stronger correlations among tests with verbal content and among tests with numeric content allowed the correlations among tests with pictorial content to become more pronounced or visible in the configuration instead of being "pulled" into CHC ability clusters. Content features may be considered when interpreting scores from these measures.

Naming Speed Literacy and Naming Speed Quantity tests were new to the WISC-V and meant to be sensitive to specific learning disorder-reading and -mathematics, respectively (Wechsler 2014). These tests measure rapid automatic naming, another component process that is important for efficient and accurate reading (Norton and Wolf 2012). Visible in Figure 2 interactive 3D MDS configuration, Naming Speed tests from the WISC-V were on the same side of the configuration as Digit Span, Letter-Number Sequencing, and Oral Reading Fluency. Each of these tests require verbal responses, but they also likely measure one or more latent Gwm narrow abilities: auditory short-term storage, visual-spatial short-term storage, attentional control, and working memory capacity (Schneider and McGrew 2018). The location of Naming Speed tests in the current study was between Gwm and Gs subtests, suggesting that they measure a blend of attentional control and processing speed (Meyer and Reynolds 2018). It is also notable that these tests share letter and number content in addition to cognitive processes.

Last, with the Wechsler data, Oral Reading Fluency was in the reading cluster, but it was located near Digit Span and Letter-Number Sequencing auditory short-term storage tests within Gwm. Working memory predicts reading fluency in children with ADHD (Swanson and Siegel 2011) and SLD (Jacobson et al. 2011). The auditory short-term storage narrow ability within working memory is one component process that contributes to reading fluency and comprehension (Norton and Wolf 2012; Schneider and McGrew 2018). Oral Reading Fluency may have been closer to the auditory short-term storage tests than Word Reading and Pseudoword Decoding because the Oral Reading Fluency task contains context and connected text around each word that an examinee must hold in immediate awareness and manipulate to comprehend what has been read already and predict what is coming next.

4.3. Regions and Fluency

McGrew et al. (2014) introduced a broader organization of tests, called regions, with MDS of the WJ IV: auditory-linguistic, figural-visual, reading-writing, quantitative-numeric regions, and speed-fluency. Auditory-linguistic, figural-visual, reading-writing, and quantitative-numeric regions were similarly visible in the WISC-V and WIAT-III map and each of the four Kaufman maps. These regions now have support from three different tests with three different samples. Some differences were found in the current study regarding speed-fluency, though these are likely due to test sampling differences.

A speed-fluency region was visible in the WISC-V and WIAT-III MDS map. The WISC-V and WIAT-III speed-fluency region included three math fluency tests, two speed of lexical access tests, and three processing speed tests. The WISC-V and WIAT-III speed-fluency region did not include Oral Reading Fluency. The Oral Reading Fluency test measures speed and accuracy in reading. The reading skills and accuracy required in Oral Reading Fluency are also required in the other word reading tests and may explain why Oral Reading Fluency was in the reading-writing region instead of closer to the speed-fluency tests. It is notable that though the math fluency tests were near the other speed-fluency tests, the math fluency tests were in the overlapping space between the quantitative-numeric and speed-fluency regions.

Different from the WISC-V and WIAT-III map speed-fluency region, in each Kaufman map, a retrieval fluency region was evident. This region contained oral language tests of speed of lexical access or retrieval fluency that were separate from oral language tests in the auditory-linguistic area. The KABC-II does not include processing tests. The Kaufman reading fluency tests require examinees to read real and nonreal words in isolation instead of in sentences and were located in the reading-writing region, though they were on the side of the reading-writing region closest to the speed of lexical access tests.

Locations of fluency tests in the MDS maps have implications for understanding academic difficulties related to fluency. Difficulties in reading and math fluency have been shown to co-occur, especially after second grade (Koponen et al. 2018) and it is necessary to understand whether disfluency come from a process deficit to intervene and remediate effectively. If fluency tests were located near each other in an academic fluency region of an MDS configuration of intelligence and academic achievement test scores, it would not lend causal evidence, but it would provide information about possible shared characteristics among fluency tests beyond academic content characteristics. In the WISC-V and WIAT-III MDS map and the WIAT-III only map, reading and math fluency tests were located near tests of the same academic domain. Based on results with these tests, fluency test scores should be kept with test of the same academic domain. Kaufman maps differed from WISC-V and WIAT-III regions in this study. There was not a speed-fluency region in the Kaufman maps even though Naming Facility and Associational Fluency tests from the KTEA-II are similar to the, the Naming Speed Literacy and Naming Speed Quantity tests from the WISC-V. The KABC-II does not include processing speed tests and KTEA-II reading fluency tests were closer to the reading-writing region. This meant that in the Kaufman maps, there was not a speed-fluency region. Instead, there was a narrower retrieval fluency region. The results support the split of Gr from the Glr cluster (Jewsbury and Bowden 2016). None of the MDS maps in this study included writing fluency tests, but analysis with additional academic fluency tests may result in different findings.

Additionally, academic fluency tests in the current study were not located between the higher-order thinking tests and basic skills tests as the conceptual framework from basic to higher-order academic skills would suggest (Mather and Wendling 2015). Basic skills tests and fluency tests are both simple tests in which examinees apply rules (they do not infer rules or relations between concepts). Fluency tests introduce speed to measure the automaticity with which the examinee completes the test. Though, fluency is not just speed of completing the task; accuracy matters too, so much that Paige and Magpuri-Lavell (2014) call it "accumaticity" in the context of reading fluency. Academic fluency tests were located around the periphery of the MDS map, suggesting that academic fluency tests in these assessments are simple, like the basic skill tests, and not of intermediate complexity between basic and higher-order tests.

4.4. Content and Response Process Facets

Facets organize and define an observation. In the context of assessment, multiple facets define the tests that make up intelligence and academic achievement tests. Arithmetic is a numeric test (content facet), measuring latent fluid reasoning (cognitive operation facet), and examinees respond verbally (response process). Though complexity did not adequately describe test locations in the WISC-V and WIAT-III MDS configuration, follow-up analysis with these data revealed that additional features of the tests (different facets) described the placement of tests instead, consistent with previous MDS studies with intelligence tests alone (Cohen et al. 2006; Guttman and Levy 1991).

In addition to interpreting CHC and academic clusters, the MDS maps appeared to systematically organize tests by shared content. Arithmetic's highest correlations are with Math Problem Solving (.61), Math Fluency Subtraction (.59), Letter-Number Sequencing (.57), Digit Span (.50), and Numerical Operations (.46). These tests and the math fluency tests were in the center (and upper portion of the static view) in Figure 7, grouped by their numeric content. Complex reading, oral language, writing, and Gc tests were not in the

center with Arithmetic and Math Problem Solving because they were grouped by verbal content (on the left side of the static view) in Figure 7. Tests with figural/pictorial content were on the other side of the tests with numeric content (lower right of the static view) in Figure 7. Tests were organized by the content facet (Cohen et al. 2006; Guttman and Levy 1991).

MDS maps also appeared to organize tests in space by response mode processes. Tests with paper-pencil responses were grouped together (near the top of the static view) in Figure 8 and tests with verbal responses were grouped together (lower left of the static view) in Figure 8. Tests that allow verbal or manual response were near the one test (Block Design) that requires a manual response (lower right of the statice view) in Figure 8. When looking at test content and response process MDS maps together, it is notable that on the WISC-V and WIAT-III most of the verbal content tests also elicit verbal responses. Tests with other types of content (numeric; figural; letter, number, color, object) were mixed in requiring verbal, manual, or paper-pencil responses.

The AERA, APA, and NCME Standards for Educational and Psychological Testing (American Educational Research Association et al. 2014) include test content and response processes as sources of validity evidence in addition to internal structure, relations to other variables, and consequences of testing. This study supported content (Figure 7) and response process (Figure 8) validity. Previous MDS studies with Wechsler tests have demonstrated content and response facets in three dimensions (Guttman and Levy 1991) and two dimensions (Cohen et al. 2006); however, this is the first study to support content and response facets with the inclusion of academic achievement tests. Content and response process should be something that test users consider in interpretation of scores.

4.5. Limitations

This research is not without limitations. There are a few limitations regarding interpretation of the results. First, there is subjectivity in selecting the number of dimensions and in interpreting the resulting configurations. There is guidance about making an informed choice for the number of dimensions in terms of absolute fit; however, due to the limited research using MDS compared to other multivariate procedures, there is little precedent about appropriate model decisions such as number of dimensions. Further, more dimensions may exist, but it is difficult for humans to interpret findings beyond three dimensions (Ambinder et al. 2009). Visualizing and interpreting three-dimensional representations is difficult because visual information is first recorded on the retina in two dimensions and then depth is integrated (Finlayson et al. 2017). Additionally, there are capacity limits to visual working memory (Qian and Zhang 2019). There is also subjectivity in the interpretations, but the visual exploration of results also creates opportunities to examine additional facets or dimensions, like the response processes in Figure 8.

Another limitation is related to the analysis of complexity. Calculating the center of the configurations via the mean is sensitive to outliers (like the Cancellation test that is far from all other tests). The center point that represents the mean of each dimension does not necessarily represent the center of the densest part of the configuration.

Next, there were limitations related to the generalization of the results. United States demographics have changed since data were collected in the norming procedures for the measures, so the findings may not generalize to the current population. Additionally, the sample represents the English-speaking U.S. population, so the findings cannot be generalized to those whose first language is not English (see Ortiz et al. 2018).

4.6. Future Research

In the future, it would be informative to include multiple intelligence or academic achievement test batteries or narrower measures. If CHC ability clusters and academic content clusters could be more balanced (e.g., each CHC ability and academic area is measured with each response mode), correlations sampled more ability areas (i.e., more CHC narrow abilities), or included more variety within clusters (e.g., four ways to measure

Gv) cross-battery comparisons would be more accurate and implications could extend beyond the boundaries of a given test battery (Beauducel and Kersting 2002; Süß and Beauducel 2015). More balance among measures could also help to eliminate alternative explanations for some findings in this research, such as having a disproportionate number of reading tests in the analysis, which may affect the findings.

MDS will also be useful as an exploratory technique when abilities are considered for addition to the CHC taxonomy (such as Gei or Emotional Intelligence per Schneider and McGrew 2018) or to test new theoretical frameworks, such as process overlap theory to explain *g* (Conway et al. 2021). Given its unrestricted nature, it may help some with overly restricted thinking.

There is also a need to replicate with non-standardization samples (Graves et al. 2020). MDS could be useful with data from culturally and linguistically diverse students. Verbal content and response processes that may increase complexity or change task demands depending on an individual's language skills. For example, the working memory task demands of a reading comprehension test may be higher if an individual is focusing on word- or sentence-level understanding (Acosta Caballero 2012). Given that the regions (e.g., auditory-linguistic) have emerged across multiple tests across multiple samples, it would be interesting to investigate if these findings are invariant across non-representative samples.

CHC theory is a flexible taxonomy and analysis of constructs that are being considered for inclusion is another direction for future research. Broader regions, like McGrew et al. (2014) auditory-linguistic, reading-writing, quantitative-numeric, and speed-fluency regions that were supported in the current study should also be studied more with additional assessments. Emotional intelligence and sensory domains (tactile, kinesthetic, etc.) are not typically included in comprehensive psychoeducational evaluations. These and other areas of abilities should be analyzed in future studies for a more complete model of individual differences that are relevant in school, occupations, and creative pursuits.

Finally, it would be worthwhile to apply MDS analysis of variables representing individual differences across a larger time span than a few weeks between test administrations. It would be interesting to also develop a method for representing longitudinal variables in continuous geometric space.

4.7. Implications

The results from this research suggest theoretical and practical implications. Practitioners need to be aware of shared content and response processes across tests that convey similarities beyond cognitive processes. Test content and response processes are discussed in test manuals, but it is not clear how often they are considered in practice. One interpretive practice that is not supported by this research is a fluency score that cuts across academic areas. More information is needed about how fluency relates to memory processes, basic skills acquisition, and higher-order academic achievement skills (Gerst et al. 2021; Lovett et al. 2020).

5. Conclusions

Snow et al. (1984, p. 47) called MDS maps of intelligence and academic achievement correlations "The Topography of Ability and Learning Correlations." MDS configurations give a sense of the landscape that is at times obscured by the viewfinder of other multivariate analyses, like factor analysis.

Several important findings emerged from this MDS research. First, test organization by CHC ability factors and academic achievement domains was supported, and in that respect, findings were consistent with findings from factor analysis (Carroll 1993). CHC theory was a useful way to describe the findings. Second, in addition to organization by CHC ability factors and academic domains, broader regions were visible, supporting McGrew et al. (2014) findings. Third, content and response process facets are useful in understanding intelligence tests and achievement test score correlations. Practitioners need

to be aware of how test information is presented to examinees and how their responses are elicited. Last, academic fluency tests were not as distinctly or consistently located in a speed-fluency region in the test batteries examined in this study because they were near academic domains.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jintelligence10040117/s1. File S1: Multidimensional Scaling of Cognitive Ability and Academic Achievement: Supplemental Figures.

Author Contributions: Conceptualization, E.M.M. and M.R.R.; methodology, E.M.M. and M.R.R.; formal analysis, E.M.M., writing—original draft preparation, E.M.M. and M.R.R., writing—review and editing, E.M.M. and M.R.R. 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: The data presented in this study are not publicly available because they are confidential and proprietary.

Acknowledgments: The authors would like to thank Pearson for permission to use standardization data.

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

References

Acosta Caballero, Karen. 2012. The Reading Comprehension Strategies of Second Langauge Learners: A Spanish-English Study. Ph.D. thesis, University of Kansas, Kansas City, KS, USA.

Adler, Daniel, and Duncan Murdoch. 2021. rgl: 3D Visualization Using OpenGL. Version 0.105.22. Available online: https://CRAN.R-project.org/package=rgl (accessed on 2 November 2021).

Ambinder, Michael, Ranxiao Wang, James Crowell, George Francis, and Peter Brinkmann. 2009. Human Four-Dimensional Spatial Intuition in Virtual Reality. *Psychonomic Bulletin & Review* 16: 818–23.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, and Psychological Testing. 2014. Standards for Educational and Psychological Testing. American Educational Research Association. Available online: https://books.google.com/books?id=clI_mAEACAAJ (accessed on 23 January 2021).

Beauducel, André, and Martin Kersting. 2002. Fluid and Crystallized Intelligence and the Berlin Model of Intelligence Structure (Bis). *European Journal of Psychological Assessment* 18: 97–112. [CrossRef]

Benson, Nicholas, Randy Floyd, John Kranzler, Tanya Eckert, Sarah Fefer, and Grant Morgan. 2019. Test Use and Assessment Practices of School Psychologists in the United States: Findings from the 2017 National Survey. *Journal of School Psychology* 72: 29–48. [CrossRef]

Benson, Nicholas, John Kranzler, and Randy Floyd. 2016. Examining the Integrity of Measurement of Cognitive Abilities in the Prediction of Achievement: Comparisons and Contrasts across Variables from Higher-Order and Bifactor Models. *Journal of School Psychology* 58: 1–19. [CrossRef] [PubMed]

Borg, Ingwer, and Patrick Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications*, 2nd ed. Springer Series in Statistics; Berlin and Heidelberg: Springer.

Borg, Ingwer, Patrick Groenen, and Patrick Mair. 2018. *Applied Multidimensional Scaling and Unfolding*, 2nd ed. Springer Briefs in Statistics. Berlin and Heidelberg: Springer.

Breaux, Kristina. 2009. *Wechsler Individual Achievement Test*, 3rd ed. Technical Manual. Chicago: NCS Pearson.

Caemmerer, Jacqueline, Timothy Keith, and Matthew Reynolds. 2020. Beyond Individual Intelligence Tests: Application of Cattell-Horn-Carroll Theory. *Intelligence* 79: 101433. [CrossRef]

Caemmerer, Jacqueline, Danika Maddocks, Timothy Keith, and Matthew Reynolds. 2018. Effects of Cognitive Abilities on Child and Youth Academic Achievement: Evidence from the WISC-V and WIAT-III. *Intelligence* 68: 6–20. [CrossRef]

Carroll, John. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge: Cambridge University Press.

Cattell, Raymond. 1943. The Measurement of Adult Intelligence. *Psychological Bulletin* 40: 153–93. [CrossRef]

Cohen, Arie, Catherine Fiorello, and Frank Farley. 2006. The Cylindrical Structure of the Wechsler Intelligence Scale for Children-IV: A Retest of the Guttman Model of Intelligence. *Intelligence* 34: 587–91. [CrossRef]

Conway, Andrew, Kristof Kovacs, Han Hao, Kevin P. Rosales, and Jean-Paul Snijder. 2021. Individual Differences in Attention and Intelligence: A United Cognitive/Psychometric Approach. *Journal of Intelligence* 9: 34. [CrossRef]

Cormier, Damien, Okan Bulut, Kevin McGrew, and Jessica Frison. 2016. The Role of Cattell-Horn-Carroll (Chc) Cognitive Abilities in Predicting Writing Achievement During the School-Age Years. *Psychology in the Schools* 53: 787–803. [CrossRef]

Cormier, Damien, Kevin McGrew, Okan Bulut, and Allyson Funamoto. 2017. Revisiting the Relations between the Wj-Iv Measures of Cattell-Horn-Carroll (Chc) Cognitive Abilities and Reading Achievement During the School-Age Years. *Journal of Psychoeducational Assessment* 35: 731–54. [CrossRef]

Finlayson, Nonie, Xiaoli Zhang, and Julie Golomb. 2017. Differential Patterns of 2d Location Versus Depth Decoding Along the Visual Hierarchy. *NeuroImage* 147: 507–16. [CrossRef] [PubMed]

Floyd, Randy, Jeffrey Evans, and Kevin McGrew. 2003. Relations between Measures of Cattell-Horn-Carroll (Chc) Cognitive Abilities and Mathematics Achievement across the School-Age Years. *Psychology in the Schools* 40: 155–71. [CrossRef]

Floyd, Randy, Timothy Keith, Gordon Taub, and Kevin McGrew. 2007. Cattell-Horn-Carroll Cognitive Abilities and Their Effects on Reading Decoding Skills: G Has Indirect Effects, More Specific Abilities Have Direct Effects. *School Psychology Quarterly* 22: 200–33. [CrossRef]

Garcia, Gailyn, and Mary Stafford. 2000. Prediction of Reading by Ga and Gc Specific Cognitive Abilities for Low-Ses White and Hispanic English-Speaking Children. *Psychology in the Schools* 37: 227–35. [CrossRef]

Gerst, Elyssa, Paul Cirino, Kelly Macdonald, Jeremy Miciak, Hanako Yoshida, Steven Woods, and Cullen Gibbs. 2021. The Structure of Processing Speed in Children and Its Impact on Reading. *Journal of Cognition and Development* 22: 84–107. [CrossRef] [PubMed]

Graham, John. 2009. Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology* 60: 549–76. [CrossRef] [PubMed]

Graves, Scott, Leanne Smith, and Kayla Nichols. 2020. Is the WISC-V a Fair Test for Black Children: Factor Structure in an Urban Public School Sample. *Contemporary School Psychology* 25: 157–69. [CrossRef]

Gustafsson, Jan-Eric. 1984. A Unifying Model for the Structure of Intellectual Abilities. *Intelligence* 8: 179–203. [CrossRef]

Gustafsson, Jan-Eric, and Gudrun Balke. 1993. General and Specific Abilities as Predictors of School Achievement. *Multivariate Behavioral Research* 28: 407–34. [CrossRef]

Guttman, Louis, and Shlomit Levy. 1991. Two Structural Laws for Intelligence Tests. *Intelligence* 15: 79–103. [CrossRef]

Hajovsky, Daniel, Ethan Villeneuve, Joel Schneider, and Jacqueline Caemmerer. 2020. An Alternative Approach to Cognitive and Achievement Relations Research: An Introduction to Quantile Regression. *Journal of Pediatric Neuropsychology* 6: 83–95. [CrossRef]

Hajovsky, Daniel, Matthew Reynolds, Randy Floyd, Joshua Turek, and Timothy Keith. 2014. A Multigroup Investigation of Latent Cognitive Abilities and Reading Achievement Relations. *School Psychology Review* 43: 385–406. [CrossRef]

Jacobson, Lisa, Matthew Ryan, Rebecca Martin, Joshua Ewen, Stewart Mostofsky, Martha Denckla, and Mark Mahone. 2011. Working Memory Influences Processing Speed and Reading Fluency in Adhd. *Child Neuropsychology* 17: 209–24. [CrossRef] [PubMed]

Jensen, Aarthur. 1998. *The G Factor: The Science of Mental Ability*. Westport: Praeger Publishing.

Jewsbury, Paul, and Stephen Bowden. 2016. Construct Validity of Fluency and Implications for the Factorial Structure of Memory. *Journal of Psychoeducational Assessment* 35: 460–81. [CrossRef]

Kaufman, Alan, and Nadeen Kaufman. 2004. *Kaufman Assessment Battery for Children—Second Edition (KABC-II) Manual*. Circle Pines: American Guidance Service.

Kaufman, Alan. 2004. *Kaufman Test of Educational Achievement, Second Edition Comprehensive Form Manual*. Circle Pines: American Guidance Service.

Kaufman, Alan, Nadeen Kaufman, and Kristina Breaux. 2014. *Technical & Interpretive Manual. Kaufman Test of Educational Achievement—Third Edition*. Bloomington: NCS Pearson.

Kaufman, Scott, Matthew Reynolds, Xin Liu, Alan Kaufman, and Kevin McGrew. 2012. Are Cognitive G and Academic Achievement G One and the Same G? An Exploration on the Woodcock-Johnson and Kaufman Tests. *Intelligence* 40: 123–38. [CrossRef]

Keith, Timothy. 1999. Effects of General and Specific Abilities on Student Achievement: Similarities and Differences across Ethnic Groups. *School Psychology Quarterly* 14: 239–62. [CrossRef]

Keith, Timothy, and Matthew Reynolds. 2010. Cattell-Horn-Carroll Abilities and Cognitive Tests: What We've Learned from 20 Years of Research. *Psychology in the Schools* 47: 635–50. [CrossRef]

Koponen, Tuire, Mikko Aro, Anna-Maija Poikkeus, Pekka Niemi, Marja-Kristilina Lerkkanen, Timo Ahonen, and Jari-Erik Nurmi. 2018. Comorbid Fluency Difficulties in Reading and Math: Longitudinal Stability across Early Grades. *Exceptional Children* 84: 298–311. [CrossRef]

Kruskal, Joseph. 1964. Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrika* 29: 115–29. [CrossRef]

Li, Xueming, and Stephen Sireci. 2013. A New Method for Analyzing Content Validity Data Using Multidimensional Scaling. *Educational and Psychological Measurement* 73: 365–85. [CrossRef]

Lichtenberger, Elizabeth, and Kristina Breaux. 2010. *Essentials of WIAT-III and KTEA-II Assessment*. Hoboken: John Wiley & Sons.

Lohman, David, and Joni Lakin. 2011. Intelligence and Reasoning. In *The Cambridge Handbook of Intelligence*. Edited by Robert Sternberg and Scott Kaufman. Cambridge: Cambridge University Press, pp. 419–41.

Lovett, Benjamin, Allyson Harrison, and Irene Armstrong. 2020. Processing Speed and Timed Academic Skills in Children with Learning Problems. *Applied Neuropsychology Child* 11: 320–27. [CrossRef] [PubMed]

Mair, Patrick, Ingwer Borg, and Thomas Rusch. 2016. Goodness-of-Fit Assessment in Multidimensional Scaling and Unfolding. *Multivariate Behavioral Research* 51: 772–89. [CrossRef] [PubMed]

Mair, Patric, Jan de Leeuw, Patrick Groenen, and Ingwer Borg. 2021. Smacof R Package Version 2.1-2. Available online: https://cran.r-project.org/web/packages/smacof/smacof.pdf (accessed on 28 September 2022).

Marshalek, Brachia, David Lohman, and Richard Snow. 1983. The Complexity Continuum in the Radex and Hierarchical Models of Intelligence. *Intelligence* 7: 107–27. [CrossRef]

Mather, Nancy, and Bashir Abu-Hamour. 2013. Individual Assessment of Academic Achievement. In *Apa Handbook of Testing and Assessment in Psychology: Testing and Assessment in School Psychology and Education*. Edited by Kurt Geisinger. Washington: American Psychological Association, pp. 101–28.

Mather, Nancy, and Barbara Wendling. 2015. *Essentials of Wj Iv Tests of Achievement*. Hoboken: John Wiley & Sons.

McGill, Ryan. 2020. An Instrument in Search of a Theory: Structural Validity of the Kaufman Assessment Battery for Children—Second Edition Normative Update at School-Age. *Psychology in the Schools* 57: 247–64. [CrossRef]

McGrew, Kevin. 2012. *Implications of 20 Years of Chc Cognitive-Achievement Research: Back-to-the-Future and Beyond*. Medford: Richard Woodcock Institute, Tufts University.

McGrew, Kevin, Erica LaForte, and Fredrick Schrank. 2014. *Technical Manual. Woodcock-Johnson Iv*. Rolling Meadows: Riverside.

McGrew, Kevin, and Barbara Wendling. 2010. Cattell-Horn-Carroll Cognitive-Achievement Relations: What We Have Learned from the Past 20 Years of Research. *Psychology in the Schools* 47: 651–75. [CrossRef]

McGrew, Kevin. 2009. CHC Theory and the Human Cognitive Abilities Project: Standing on the Shoulders of the Giants of Psychometric Intelligence Research. *Intelligence* 37: 1–10. [CrossRef]

Meyer, Emily, and Matthew Reynolds. 2018. Scores in Space: Multidimensional Scaling of the WISC-V. *Journal of Psychoeducational Assessment* 36: 562–75. [CrossRef]

Niileksela, Christopher, and Matthew Reynolds. 2014. Global, Broad, or Specific Cognitive Differences? Using a Mimic Model to Examine Differences in Chc Abilities in Children with Learning Disabilities. *Journal of Learning Disabilities* 47: 224–36. [CrossRef]

Niileksela, Christopher, Matthew Reynolds, Timothy Keith, and Kevin McGrew. 2016. A Special Validity Study of the Woodcock-Johnson IV: Acting on Evidence for Specific Abilities. In *WJ IV Clinical Use and Interpretation: Scientist-Practitioner Perspectives (Practical Resources for the Mental Health Professional)*. Edited by Dawn P. Flanagan and Vincent C. Alfonso. San Diego: Elsevier, chp. 3. pp. 65–106.

Norton, Elizabeth, and Maryanne Wolf. 2012. Rapid Automatized Naming (Ran) and Reading Fluency: Implications for Understanding and Treatment of Reading Disabilities. *Annual Review of Psychology* 63: 427–52. [CrossRef]

Ortiz, Samuel, Nicole Piazza, SSalvador Hector Ochoa, and Agnieszka Dynda. 2018. Testing with Culturally and Linguistically Diverse Populations: New Directions in Fairness and Validity. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. Edited by Dawn Flanagan and Erin McDonough. New York: Guilford Publications, pp. 684–712.

Paige, David, and Theresa Magpuri-Lavell. 2014. Reading Fluency in the Middle and Secondary Grades. *International Electronic Journal of Elementary Education* 7: 83–96.

Qian, Jiehui, and Ke Zhang. 2019. Working Memory for Stereoscopic Depth Is Limited and Imprecise—Evidence from a Change Detection Task. *Psychonomic Bulletin & Review* 26: 1657–65.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rabin, Laura, Emily Paolillo, and William Barr. 2016. Stability in Test-Usage Practices of Clinical Neuropsychologists in the United States and Canada over a 10-Year Period: A Follow-up Survey of Ins and Nan Members. *Archives of Clinical Neuropsychology* 31: 206–30. [CrossRef] [PubMed]

Reynolds, Matthew, Timothy Keith, Jodene Goldenring Fine, Melissa Fisher, and Justin Low. 2007. Confirmatory Factor Structure of the Kaufman Assessment Battery for Children—Second Edition: Consistency with Cattell-Horn-Carroll Theory. *School Psychology Quarterly* 22: 511–39. [CrossRef]

Reynolds, Matthew, and Timothy Keith. 2007. Spearman's Law of Diminishing Returns in Hierarchical Models of Intelligence for Children and Adolescents. *Intelligence* 35: 267–81. [CrossRef]

Reynolds, Matthew, and Timothy Keith. 2017. Multi-Group and Hierarchical Confirmatory Factor Analysis of the Wechsler Intelligence Scale for Children-Fifth Edition: What Does It Measure? *Intelligence* 62: 31–47. [CrossRef]

Schneider, Joel, and Kevin McGrew. 2018. The Cattell-Horn-Carroll Theory of Cognitive Abilities. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues*. Edited by Dawn Flanagan and Erin McDonough. New York: Guilford Press, pp. 73–164.

Snow, Richard, Patrick Kyllonen, and Brachia Marshalek. 1984. The Topography of Learning and Ability Correlations. *Advances in the Psychology of Human Intelligence* 2: 47–103.

Süß, Heinz-Martin, and André Beauducel. 2015. Modeling the Construct Validity of the Berlin Intelligence Structure Model. *Estudos de Psicologia* 32: 13–25. [CrossRef]

Swanson, Lee, and Linda Siegel. 2011. Learning Disabilities as a Working Memory Deficit. *Experimental Psychology* 49: 5–28.

Thurstone, Louis. 1938. Primary mental abilities. *Psychometric Monographs* 1: 270–75.

Tucker-Drob, Elliot, and Timothy Salthouse. 2009. Confirmatory Factor Analysis and Multidimensional Scaling for Construct Validation of Cognitive Abilities. *International Journal of Behavioral Development* 33: 277–85. [CrossRef] [PubMed]

Vanderwood, Michael, Kevin McGrew, Dawn Flanagan, and Timothy Keith. 2002. The Contribution of General and Specific Cognitive Abilities to Reading Achievement. *Learning and Individual Differences* 13: 159–88. [CrossRef]

Vernon, Philip E. 1950. *The Structure of Human Abilities*. London: Methuen. [CrossRef]
Wechsler, David. 2014. *Technical and Interpretive Manual for the Wechsler Intelligence Scale for Children–Fifth Edition (WISC-V)*. Bloomington: Pearson.
Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis Version*. Abingdon: Taylor & Francis.

Journal of
Intelligence

MDPI

Article

The Influence of Culture Capital, Social Security, and Living Conditions on Children's Cognitive Ability: Evidence from 2018 China Family Panel Studies

Xianhua Dai [1,2,*] and Wenchao Li [1]

[1] School of Public Administration, Central China Normal University, Wuhan 430079, China; liwenchao@mails.ccnu.edu.cn
[2] Center for Labor and Social Security Research, Central China Normal University, Wuhan 430079, China
* Correspondence: xhdai@ccnu.edu.cn

Abstract: The aim of this study was to analyze the influence of economic capital, culture capital, social capital, social security, and living conditions on children's cognitive ability. However, most studies only focus on the impact of family socio-economic status/culture capital on children's cognitive ability by ordinary least squares regression analysis. To this end, we used the data from the China Family Panel Studies in 2018 and applied proxy variable, instrumental variables, and two-stage least squares regression analysis with a total of 2647 samples with ages from 6 to 16. The results showed that family education, education expectation, books, education participation, social communication, and tap water had a positive impact on both the Chinese and math cognitive ability of children, while children's age, gender, and family size had a negative impact on cognitive ability, and the impact of genes was attenuated by family capital. In addition, these results are robust, and the heterogeneity was found for gender and urban location. Specifically, in terms of gender, the culture, social capital, and social security are more sensitive to the cognitive ability of girls, while living conditions are more sensitive to the cognitive ability of boys. In urban locations, the culture and social capital are more sensitive to rural children's cognitive ability, while the social security and living conditions are more sensitive to urban children's cognitive ability. These findings provide theoretical support to further narrow the cognitive differences between children from many aspects, which allows social security and living conditions to be valued.

Keywords: culture capital; economic capital; social capital; social security; living conditions; cognitive ability; heterogeneity

Citation: Dai, Xianhua, and Wenchao Li. 2022. The Influence of Culture Capital, Social Security, and Living Conditions on Children's Cognitive Ability: Evidence from 2018 China Family Panel Studies. *Journal of Intelligence* 10: 19. https://doi.org/10.3390/jintelligence10020019

Received: 24 January 2022
Accepted: 23 March 2022
Published: 25 March 2022

1. Introduction

This study analyzed the influence of economic capital, culture capital, social capital, social security, and living conditions on children's cognitive ability.

With deepening development and reformation of education, the human capital cultivation of children is becoming a key step for many families. A fundamental aspect of the cultivation is children's cognitive ability, which is the ability of human beings to extract, store, and use information from the objective world. It mainly involves human abstract thinking, logical deduction, and memory (Autor 2014). As documented, there is a significant correlation between family factors and children's cognitive ability (Zimmer et al. 2007; Kleinjans 2010; Li 2012, 2017; Saasa 2018; Fan et al. 2019; Wang and Lin 2021). Specifically, there are three different capital theories that focus on the impact of family on children's cognitive ability, namely economic capital, cultural capital, and social capital (Bourdieu and Wacquant 1992; Farkas 2003). In particular, the impact of cultural capital is particularly important (Li and Zhao 2017; Yao and Ye 2018; Zhang and Su 2018; Hong and Zhang 2021) since economic capital reflects its value only by cultural capital (Hong and Zhao 2014). In addition, there are great differences in economic capital, social capital, and cultural capital

between urban and rural families, which leads to the urban–rural education gap (Jin 2019). These findings concentrate on the influence of family capital on children's cognitive ability, but the social security and living conditions are not touched upon. In contrast, this study investigated the influence of all those factors on children's cognitive ability, in particular, the social security and living conditions.

Different family capital has corresponding measurement indicators. In particular, economic capital includes family income (Yang and Wan 2015; Fang and Hou 2019; Hou et al. 2020), health investment (Shen 2019; Wu et al. 2021), and education expenditure (Lin et al. 2021; Fang and Huang 2020), which refers to the sum of economic related resources owned by a family (Xue and Cao 2004). Culture capital is not only reflected in the diplomas obtained by family members, but also in the educational concept, attitude, and expectation of parents for their children (Guo and Min 2006), which includes three forms: concrete culture capital, such as family parenting (Zhang et al. 2017; Huang 2018), lifestyle (Wu et al. 2020), education expectation (Gu and Yang 2013; Wang and Shi 2014; Xue 2018; Zhou et al. 2019), participation (Wei et al. 2015; Liu et al. 2015; Liang et al. 2018); objectified culture capital, including books (Hong and Zhao 2014; Yan 2017); and institutionalized culture capital, referring to the educational diploma obtained (Xie and Xie 2019; Zhu et al. 2018). From the perspective of micro social network, social capital referred to in this paper is defined as a kind of resource embedded in the network (Granovetter 1973), which takes social capital as a new form of capital, so that actors can obtain a better professional position or business opportunities, so as to affect the income return (Lin 2005). In specific, social capital includes occupation (James 2000; Teacherman 2000; Fang and Feng 2005; Zhou and Zou 2016; Zhu and Zhang 2020), social communication (Putnam 2000; Liang 2020; Yang and Zhang 2020), information utilization (Cao et al. 2018; Zheng et al. 2021), and human expenditure (Wang and Gong 2020).

Social security can improve residents' household consumption (Fang and Zhang 2013; Yang and Yuan 2019) and alleviate economic poverty (Guo and Sun 2019) through income redistribution, which can increase the economic capital of families and affect investment in children. Thus, the social security affects children's cognitive ability, including medical insurance (Chen et al. 2020), endowment insurance (Xue et al. 2021), and government support (Liu and Xue 2021; Yin and Fan 2021). Living conditions refer to the family infrastructure and facilities that affect children's lives, including safe drinking water, sanitary toilets, clean energy, waste treatment, and sewage treatment (Zhao et al. 2018). In particular, exposure to air pollution (Chen et al. 2017a; Schikowski and Altug 2020; Nauze and Severnini 2021), water (Chen et al. 2017b; Gao et al. 2021), and fuel (Cong et al. 2021; Chen et al. 2021) also affects cognitive ability. Other factors include family structure (Zhang 2020; Jiang and Zhang 2020), family size (Liu and Jin 2020; Fang et al. 2020), and family health (Li and Fang 2019). Unlike previous work, this study applied instrumental variables and two-stage least squares regression analysis to solve the endogenous problem, assessing the influence of numerous factors on children's cognitive ability. The robustness of this study's results was assessed by controlling sample size and increasing variables.

In addition, children's individual and social characteristics affect cognitive ability. For example, the performance of girls is better than that of boys, although the gender difference is decreasing (Hao 2018). The older the migrant child, the worse the academic performance (Wang and Chu 2019). Number of siblings has a significant impact on youth's cognitive ability (Tao 2019). In contrast, this study investigated heterogeneity in gender and urban location for those influences.

This study examined the impact of numerous factors, including social security and living conditions, on children's cognitive ability, using data from the China Family Panel Studies in 2018. Rather than the ordinary least squares method, the study used two-stage least squares regression to solve endogeneity. In addition, we explored heterogeneity in gender and urban location and the impact of those factors on children's cognitive ability. These results obtained may provide guidance for the government, society, and families to improve children's cognitive ability.

The remainder of this paper is organized as follows. Section 2 describes data, variables, and summary statistics. Section 3 outlines the basic model for the influence of those factors on children's cognitive ability. Section 4 describes the instrumental variable test, endogeneity test, empirical results, and robustness test. Section 5 outlines the heterogeneity analysis of gender and urban location. Section 6 concludes.

2. Data, Variable, and Summary Statistics

2.1. Data

This study used the data from the China Family Panel Studies (CFPS), a tracking survey of individuals, families, and communities implemented by China Social Science Investigation Center of Peking University, which aims to reflect the changes of China's society, economy, education, and health. The data sample covers 25 provinces/cities/autonomous regions, and the respondents include all family members. In the implementation of the survey, the multi-stage, implicit stratified, and population scale proportional sampling method was used. The main research object of this study was children aged 6–16. Since the respondents of the CFPS personal self-administered questionnaire are children over nine years old, and children's cognition of their own situation is not necessarily accurate, this study mainly used the children's proxy questionnaire and combined the relevant variables such as parents' situation in the personal self-administered questionnaire and family basic information in the family questionnaire. The data supported this work. The basic information related to families, parents, and their children in 2018 was extracted and matched with the data.

2.2. Explained Variables

Following Li and Shen (2021), Wu et al. (2020), and Dong and Zhou (2019), children's Chinese and math scores were used in this study to measure Chinese cognitive understanding ability and math reasoning cognitive ability, respectively, using the "How about Chinese score" and "How about math score" tests in the CFPS questionnaire, both of which use ordinal categorical variables (1 for "fail", 2 for "intermediate", 3 for "good", and 4 for "distinction").

2.3. Explanatory Variables

In this study, the main explanatory variables were divided into five parts. They are economic capital, culture capital, social capital, social security, and living conditions.

Economic capital was measured by the family income, children's health investment, and education investment. They are all continuous variables and were added 1 before taking the natural logarithm.

Culture capital was measured by the questions of "How many books do you have in your family?", "What is the highest degree you have completed?", "What level of education do you want your child to attain?", "How often do you discuss what's happening at school with your child?", and "When your children's grades are not satisfactory, which way do you usually deal with them?". They represent the family books, education, educational expectation, educational participation, and parenting style, respectively. There are three aspects of culture capital, namely the objective, institutional, and concrete culture capital (Bourdieu and Passeron 1977). For family education and education expectation, 0 is for illiterate/semi-illiterate, 1 for nursery, 2 for kindergarten, 3 for primary school, 4 for junior middle school, 5 for senior middle school, 6 for junior college, 7 for undergraduate, 8 for master, and 9 for doctor. For parenting style, we redefined scolding the child, spanking the child, and restricting the child's activities as 0, and contacting the teacher, telling the child to study harder, helping the child more, and doing nothing as 1. Among them, 0 is for stern parenting, and 1 is for gentle parenting. Family books and children's education participation are continuous variables, and the number of books was added 1 before taking the natural logarithm. In addition, family lifestyle consists of smoking, drinking, exercise, and lunch break, which is an ordered variable.

Social capital was measured by "nature of work", "information utilization", "social communication", and "human expenditure". For job, 1 is unemployed, 2 is agricultural work, and 3 is non-agricultural work. We used the questions of "Do you use a mobile phone?", "Do you use mobile devices?", and "Do you use a computer to surf the Internet?" to measure the information utilization. We defined information utilization as follows: 0 means that none is used, 1 means that at least one is used, 2 means that at least two are used, and 3 means that at least three are used. The questions of "How good do you think your relationship is?" and "How do you rate your trust in your neighbors?" were used to measure the social communication. We summed and then averaged the answers to these two questions and obtained a continuous variable. Human expenditure is a continuous variable and was added 1 before taking the natural logarithm.

Social security was measured by the participation of medical and endowment insurance and government support. Among them, medical and endowment insurance are continuous variables. For government support, 0 is for not accepting subsidies, and 1 is for accepting the subsidies.

Living conditions were measured by the questions about "water for cooking", "cooking fuel", and "indoor air purification", and the answer 0 is for no and 1 is for yes. Specifically, for tap water, 0 represented no tap water use, and 1 is for tap water use. For cooking fuel, 0 is for no use of clean fuel, and 1 is for clean fuel use. For air purification, 0 is for no air purification, and 1 is for use of air purification. In addition, for gender, 0 is for women and 1 is for men. The registered residence was redefined: 0 is for rural, and 1 is for urban. The registered marital status was redefined: 0 is for unmarried, and 1 is for married. For nationality, 0 is for others, and 1 is for Han nationality. Family age, the child's age, and family size are the continuous variables. For family health, 1 denotes unhealthy, 2 relatively unhealthy, 3 average, 4 relatively healthy, and 5 very healthy. We used the question "How many times a week do you eat with your family?" to measure parenthood, which is a continuous variable.

In addition, we consider parents' cognitive ability as proxy variable of genes. According to the CFPS in 2018 for the children's questionnaire, the respondents may be father or mother. Following Li and Zhang (2018), we select two dimensions of father's or mother's word ability and mathematical ability to construct parents' cognitive ability indicators. To compare, we standardized the scores of word ability and mathematical ability, and added up to obtain a comprehensive cognitive ability, which is recorded as family cognitive ability.

Table 1 shows the summary statistics of variables.

Table 1. Summary statistics of variables.

	Number	Min (M)	Max (X)	Average (E)	Standard Error	Standard Deviation	Variance
Chinese (understanding)	2647	1	4	2.760	(0.019)	0.978	0.956
Math (reasoning)	2647	1	4	2.790	(0.020)	1.041	1.083
Child's age	2647	6	16	10.90	(0.049)	2.538	6.442
Child's gender	2647	0	1	0.540	(0.010)	0.499	0.249
Child's nationality	2647	0	1	1.000	(0.001)	0.043	0.002
Residence	2647	0	1	0.180	(0.007)	0.381	0.145
Urban–rural	2647	0	1	0.430	(0.010)	0.495	0.245
Family age	2647	18	78	41.66	(0.178)	9.181	84.288
Family gender	2647	0	1	0.350	(0.009)	0.477	0.228
Family marriage	2647	0	1	0.960	(0.004)	0.201	0.041
Family size	2647	2	15	5.260	(0.038)	1.978	3.912
Family income	2647	0	13.82	10.744	(0.021)	1.071	1.148
Family health investment	2647	0	11.37	4.304	(0.055)	2.814	7.917
Family education investment	2647	0	11.69	7.265	(0.035)	1.776	3.153
Family education	2647	0	8	3.450	(0.036)	1.834	3.363
Family books	2647	0	9	2.510	(0.038)	1.931	3.727

Table 1. *Cont.*

	Number	Min (M)	Max (X)	Average (E)	Standard Error	Standard Deviation	Variance
Family education expectation	2647	3	9	6.800	(0.019)	1.002	1.005
Family parenting	2647	0	1	0.890	(0.006)	0.312	0.098
Family education participation	2647	1	5	3.260	(0.022)	1.135	1.287
Family lifestyle	2647	0	4	1.87	(0.016)	0.802	0.643
Family occupation	2647	1	3	2.400	(0.012)	0.624	0.389
Family information	2647	0	3	1.790	(0.015)	0.753	0.566
Family human expenditure	2647	0	11.00	7.372	(0.042)	2.178	4.743
Family social communication	2647	1	10	6.830	(0.031)	1.583	2.505
Family medical insurance	2647	0	3	0.950	(0.006)	0.292	0.086
Family endowment insurance	2647	0	4	0.720	(0.011)	0.565	0.319
Family government support	2647	0	1	0.500	(0.010)	0.500	0.250
Tap water	2647	0	1	0.73	(0.009)	0.445	0.198
Fuel	2647	0	1	0.70	(0.009)	0.458	0.210
Air purification	2647	0	1	0.03	(0.003)	0.178	0.032
Family heath	2647	1	5	3.04	(0.023)	1.187	1.408
Family relationship	2647	0	7	6.20	(0.036)	0.851	3.425
Family Chinese cognitive ability	2647	0	34	18.33	(0.216)	11.121	123.676
Family math cognitive ability	2647	0	24	8.74	(0.096)	4.637	21.504
Family cognitive ability	2647	−3.53	4.70	0.00	(0.034)	1.729	2.990

By deleting invalid values, 2647 final valid samples were included. As shown in Table 1, for children's characteristics, approximately 54% of children were boys, 46% were girls, 43% lived in urban areas, 57% lived rurally, and the children's age ranged from 6 to 16. For family characteristics, approximately 35% were male, 65 were female, 96% had a spouse, the family age ranged from 18 to 78, and the average family size was 5.

For family economic capital, the mean values of family income, children's health investment, and education investment are 10.74, 4.30, and 7.27, respectively. Education investment is significantly greater than health investment. For family culture capital, approximately 89% of families adopted a mild parenting approach, the frequency of families talking with their children is 3.26, the average educational level of the family is primary school, and the family education expectation is undergraduate. The average value of family lifestyle is 1.87, indicating that families account for at least two of smoking, drinking, exercise, and lunch break. The average number collected books in the family is 2.51. Institutionalized and materialized cultural capital are not high, but the level of morphological cultural capital is relatively high, indicating that families pay more attention to education.

For family social capital, family non-agricultural employment is significantly greater than agricultural employment or unemployment; the average family information and human expenditure are 1.79 and 7.372, respectively; the popularity of social communication is 6.83; and the family social capital is moderate to good. For family social security, every family has at least one kind of medical insurance and endowment insurance, and at least half of the people have received government subsidies. For living conditions, the values for utilities of tap water, fuel, and air purification are 73%, 70%, and 3%, respectively; the popularity of tap water and clean fuel is high, while the popularity of air purifiers is low. In addition, children's Chinese and math cognitive ability were both moderate; the average cognitive ability of math is higher than that of Chinese.

For family cognitive ability, the average of Chinese and math cognitive ability is 18.33 and 8.74, respectively, and the overall level of family cognitive ability is not high. We included the standardized and aggregated comprehensive family cognitive ability in Table 1, with a maximum of 4.70 and a minimum of −3.53.

3. Basic Model

This study included 29 characteristics as covariates. To investigate effect of those factors on children's Chinese cognitive ability and math cognitive ability, respectively, we established the following model.

$$E_{ni} = \beta_0 + \sum_{k=1}^{3} \beta_{k1} C_{ki} + \sum_{j=1}^{6} \beta_{j2} F_{ji} + \sum_{l=1}^{20} \beta_{l3} S_{li} + \varepsilon_i \quad (1)$$

where E_{ni} is the n-th cognitive ability for the child i (n = 1, 2, where 1 is for Chinese and 2 for math); C_{ki} is the k-th children's characteristics for the child i (k = 1, 2, ... 3); F_{ji} is the j-th family information for the child i (j = 1, 2, ... , 6); S_{li} is the l-th family capital and family cognitive ability for the child i (l = 1, 2, ... , 20); β_{k1}, β_{j2}, and β_{l3} are the corresponding parameters to those variables, and ε_i is the regression error term.

Through the above model, we used ordinary least squares (OLS) regression to obtain results. However, due to the reverse causal relationship and confounding factor, we had to find proxy variable to genetic, instrumental variables to solve endogeneity, and verify them according to the assumptions. Thus, we used two-stage least squares (2SLS) as the main empirical approach and compared with ordinary least squares (OLS). As a robustness check, we conducted analysis by adding variables and controlling sample size. In addition, the heterogeneity in gender and urban location was checked based on two-stage least squares (2SLS).

As for the sharing genes and environment between parents and children being concerned, we make the following discussion. On the one hand, the social environment experienced by children and their parents is different. In specific, the children studied in this paper were born in the 21st century, so they did not experience major social changes and disasters. However, their parents have experienced great social changes, for example, cultural revolution, educational reform, and natural disasters. On the other hand, the inequality of family resources will lead to the inequality of children's cognitive ability and early skills dependent partly on genetics (Plomin and Stumm 2018; Silventoinen et al. 2020). Thus, these two factors usually produced an interesting phenomenon, that is, the higher the importance of one, the smaller the other. However, as resulted by Houmark et al. (2020), the relative importance of genes depends on how parents' investment is distributed among their children, whether parents or society are. As also resulted by Victor Ronda et al. (2020), the worse the childhood environment, including family resources, the weaker the role of their genes. In addition, as proved, cognitive ability can be developed through acquired cultivation (Hu and Xie 2011; Kuang et al. 2019; Zhou et al. 2021), but the cognitive ability, in this paper, refers to children's word understanding ability and mathematical reasoning ability, which are measured by the scores of Chinese and math tests, respectively, and not measured by IQ test scores, though IQ test scores largely depend on genes. Furthermore, as observed from the samples in CFPS data, Chinese and math cognitive abilities of children with the same family ID were inconsistent. In particular, since the data of the 2018 China Family Panel Studies that we applied in this work do not provide genetic information, we take parents' cognitive ability as the proxy variable of genes in regression analysis.

In this study, proxy variables meet the following two conditions: (1) After introducing proxy variables (parental cognitive ability), there is no correlation between family capital and genes. Indeed, following Zheng et al. (2018), family capital is an acquired environmental factor. (2) Once the genes are observed, parents' cognitive ability will no longer mainly explain children's cognitive ability. Specifically, parental cognitive ability is highly correlated with their genes, and parental cognitive ability is not collinear with other explanatory variables. As checked, parental cognitive ability is not related to random error, and family cognitive ability can be used as a proxy variable to reflect the genetic difference.

Following Cui and Susan (2022), instrumental variables and two stage least squares regression are applied. In particular, when the exposed group and the non-exposed group are not comparable, some background variables need be used to stratify the total group so

that the exposed sub-group and the non-exposed sub-group are comparable. Instrumental variable analysis can control those bias in observational studies (Geng 2004; Brookhart et al. 2006). The instrumental variables and two stage least squares analysis in this paper will be shown in Section 4.2.

4. Results

4.1. Results from OLS

Using the survey data of CFPS in 2018, we successively incorporated family cognitive ability and family capital into the regression and applied the ordinary least squares (OLS) method to investigate the influence of family economic capital, culture capital, social capital, social security, living conditions, and family cognitive ability on children's Chinese and math cognitive ability. After excluding the influence of collinearity, the results are shown in the second to fifth column of Table 2.

Table 2. Results for the influence of many factors on children's cognitive ability.

	Chinese (OLS) N **= 2647**	**Math (OLS)** N **= 2647**	**Chinese (OLS)** N **= 2647**	**Math (OLS)** N **= 2647**	**Chinese (2SLS)** N **= 2647**	**Math (2SLS)** N **= 2647**
Intercept term	3.736 ***	4.317 ***	1.903 ***	2.181 ***	2.968 ***	3.088 ***
	(0.458)	(0.483)	(0.511)	(0.538)	(0.641)	(0.655)
Child's age	−0.058 ***	−0.103 ***	−0.053 ***	−0.096 ***	−0.055 ***	−0.098 ***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)
Child's gender	−0.287 ***	0.000	−0.287 ***	0.001	−0.284 ***	0.004
	(0.037)	(0.039)	(0.036)	(0.038)	(0.040)	(0.041)
Child's nationality	−0.327	−0.635	−0.400	−0.698	−0.438	−0.729
	(0.426)	(0.449)	(0.416)	(0.438)	(0.459)	(0.469)
Family age	0.003	0.004*	0.005 **	0.007 ***	0.006 **	0.008 ***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
Family gender	−0.038	−0.046	−0.036	−0.067	−0.028	−0.060
	(0.039)	(0.042)	(0.041)	(0.043)	(0.045)	(0.046)
Residence	0.158 ***	0.167 ***	−0.004	−0.011	−0.129 *	−0.117
	(0.055)	(0.058)	(0.058)	(0.061)	(0.074)	(0.075)
Urban–rural	0.031	0.091**	−0.044	0.026	−0.060	0.012
	(0.042)	(0.044)	(0.044)	(0.046)	(0.049)	(0.050)
Family marriage	0.129	0.133	0.066	0.059	0.044	0.040
	(0.093)	(0.098)	(0.091)	(0.096)	(0.101)	(0.103)
Family size	−0.024 **	−0.022 **	−0.016	−0.018 *	−0.008	−0.012
	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)
Family cognitive ability	0.054 ***	0.054 ***	−0.015	−0.011	−0.022	−0.017
	(0.012)	(0.012)	(0.014)	(0.015)	(0.016)	(0.016)
Family income			−0.002	0.020	−0.002	0.023
			(0.019)	(0.020)	(0.022)	(0.023)
Children's health investment			−0.004	−0.003	−0.003	−0.003
			(0.007)	(0.007)	(0.007)	(0.007)
Children's education investment			0.014	−0.001	0.013	−0.002
			(0.011)	(0.012)	(0.012)	(0.013)
Family education			0.081 ***	0.085 ***	0.087 ***	0.090 ***
			(0.015)	(0.015)	(0.017)	(0.017)
Family education expectation			0.122 ***	0.163 ***	0.116 ***	0.158 ***
			(0.018)	(0.019)	(0.021)	(0.022)
Family books/Bookiv			0.020 *	0.019 *	0.101 **	0.089 *
			(0.010)	(0.011)	(0.046)	(0.047)
Family parenting			0.038	0.099	0.015	0.080
			(0.059)	(0.062)	(0.065)	(0.066)
Family education participation			0.082 ***	0.058 ***	0.078 ***	0.055 ***
			(0.017)	(0.018)	(0.019)	(0.020)
Family lifestyle			0.006	−0.019	−0.004	−0.026
			(0.023)	(0.025)	(0.026)	(0.027)

Table 2. *Cont.*

	Chinese (OLS) N = 2647	Math (OLS) N = 2647	Chinese (OLS) N = 2647	Math (OLS) N = 2647	Chinese (2SLS) N = 2647	Math (2SLS) N = 2647
Family occupation			−0.015	0.010	−0.014	0.011
			(0.034)	(0.035)	(0.038)	(0.038)
Family information			−0.001	0.021	−0.001	0.021
			(0.033)	(0.035)	(0.037)	(0.038)
Family human expenditure			−0.013	−0.014	−0.007	−0.009
			(0.009)	(0.009)	(0.010)	(0.011)
Family social communication			0.045 ***	0.038 ***	0.048 ***	0.039 ***
			(0.012)	(0.012)	(0.013)	(0.013)
Medical insurance/Mediv			−0.004	−0.064	−1.427 ***	−1.273 ***
			(0.065)	(0.068)	(0.466)	(0.476)
Endowment insurance			0.033	0.016	0.229 ***	0.183**
			(0.034)	(0.036)	(0.076)	(0.078)
Government support			0.014	0.008	0.043	0.033
			(0.039)	(0.041)	(0.045)	(0.045)
Tap water			0.089 **	0.058	0.091 *	0.060
			(0.043)	(0.045)	(0.048)	(0.049)
Fuel			0.048	−0.040	−0.003	−0.079
			(0.045)	(0.048)	(0.052)	(0.053)
Air purification			−0.069	0.037	−0.073	0.034
			(0.102)	(0.108)	(0.113)	(0.115)
R^2	0.062	0.081	0.121	0.139	−0.064	0.021
SER	0.949	1.000	0.921	0.971	1.014	1.035
F					30.984	30.984

Note: *, **, and *** indicate significance at 10%, 5%, and 1% level, respectively; the standard error is in brackets under the coefficient.

As shown in second and third columns of Table 2, the effect of family cognitive ability on children's cognitive ability was significant (0.054, $p < 0.01$), i.e., the shared genes partly determine children's cognitive ability. As shown in the fourth and fifth columns of Table 2, the effect of family cognitive ability is no longer significant, i.e., the role of genes will be weakened by family capital. This has also been confirmed in Victor Ronda et al. (2020). Besides, children's age (−0.053, $p < 0.01$) and gender (−0.287, $p < 0.01$) have significant influence on Chinese cognitive ability, while only children's age (−0.096, $p < 0.01$) has significant influence on math cognitive ability. The influence of children's age and gender on the two cognitive abilities are both negative, while family age (0.005, $p < 0.05$; 0.007, $p < 0.01$) has a positive effect on their children's cognitive ability for Chinese and math.

For family culture capital, family education (0.081, $p < 0.01$; 0.085, $p < 0.01$), education expectation (0.122, $p < 0.01$; 0.163; $p < 0.01$), and family books (0.020, $p < 0.1$; 0.019, $p < 0.1$) have a positive impact on the two cognitive abilities. Among them, education expectation has the greatest impact, followed by family education and family books, and the influence of education expectation and family education on math cognitive ability is greater than that of Chinese, while the influence of family books is opposite. The more frequently families participate in education (0.082, $p < 0.01$; 0.058, $p < 0.01$), the better their children's cognitive abilities, and the impact on Chinese cognitive ability is greater than the impact on math. For family social capital, the impact of social communication on both children's Chinese (0.045, $p < 0.01$) and math (0.038, $p < 0.01$) cognitive abilities is positive. For living conditions, only tap water (0.089, $p < 0.05$) exhibited a positive impact on children's Chinese cognitive ability. In general, cultural capital has the greatest impact, followed by living conditions and social capital. However, the influence of family economic capital is not significant. The above results are based on ordinary least squares (OLS).

4.2. Endogeneity Test

In Equation (1), to avoid the endogenous problems caused by omitted variables, we consider the children's characteristics and family information, including age, gender, nationality, residence, marriage, and family size. These variables have been proved to have an impact on children's cognitive ability in previous studies. In this model, the main endogenous problems may be caused by the confounding factors and mutual causality. For example, children of high cognitive ability may have better genes than those of low cognitive ability. If children of high cognitive ability do not receive the acquired training, they are also more likely to obtain high cognitive ability, since their genes are excellent. However, as summarized by Miettinen and Cook (1981), confounding factors are independent risk factors; the distribution of confounding factors in exposed population and non-exposed population is different. So, we take family cognitive ability as poxy variable of genes.

Family books and family medical insurance passed the test of endogenous variables, while the family cognitive ability did not. Possible causes are confounding factors or mutual causality. For mutual causality, family books and family medical insurance may affect children's cognitive ability. Conversely, children of higher cognitive ability may have more books bought for them by their parents to support and encourage them, and the medical insurance decision will also change (Zhang and Li 2021). Therefore, we solve these problems by selecting appropriate instrumental variables. Specifically, we adopted instrumental variables (IVs) and two-stage least squares (2SLS). We used the lag variable Bookiv as the instrumental variable of family books and the average participation rate of medical insurance (Mediv) in 28 provinces as the instrumental variable of medical insurance.

Our instrumental variables satisfy the assumptions of IVs (Angrist et al. 1996). Specifically, Bookiv is highly correlated with family books, and its impact on children's cognitive ability is realized through family books, rather than directly affecting children's cognitive ability. For Mediv, which is highly correlated with family medical insurance, the average participation rate does not have a direct impact on children's cognitive ability. No other confounding factors exist between instrumental variables and children's cognitive ability. In the previous literature, the factors that affect children's cognitive ability were included in the regression to avoid the influence of confounding factors. To ensure that the IV estimation was reliable, we used the weak instrumental variable test, and as the result show, family books and medical insurance are endogenous variables. Furthermore, the Cragg–Donald–Wald F is 30.984, which is obviously greater than 10.

As shown in sixth and seventh columns of Table 2, children's age (-0.055, $p < 0.01$) and gender (-0.284, $p < 0.01$) have significant influence on their Chinese cognitive ability. The influence of children's age and gender on the two cognitive abilities is negative, while the influence of family age (0.006, $p < 0.05$; 0.008, $p < 0.01$) is positive. For family culture capital, family education (0.087, $p < 0.01$; 0.090, $p < 0.01$), education expectation (0.116, $p < 0.01$; 0.158; $p < 0.01$), and books (0.101, $p < 0.05$; 0.089, $p < 0.1$) have a positive impact on the two cognitive abilities. Similarly, education expectation has the greatest impact, followed by family education and books, and the influence of education expectation and family education on math cognitive ability is greater than that of Chinese, respectively, while the influence of family books is the opposite. The more frequently families participate in education (0.078, $p < 0.01$; 0.055, $p < 0.01$), the better their children's cognitive abilities, and the impact on Chinese cognitive ability is greater than on math. For family social capital, the impact of social communication on both children's Chinese (0.048, $p < 0.01$) and math (0.039, $p < 0.01$) cognitive abilities is positive. In addition, for family social security, medical insurance (-1.427, $p < 0.01$; -1.273, $p < 0.01$) has negative impact on both Chinese and math cognitive abilities, while endowment insurance (0.229, $p < 0.01$; 0.183, $p < 0.05$) has positive impact on both Chinese and math cognitive abilities. Tap water (0.091, $p < 0.1$) has a positive impact on children's Chinese cognitive ability. After introducing instrumental variables, the impact of family books and medical insurance on children's cognitive ability increased. The above results are based on the two-stage least squares (2SLS).

4.3. Robustness Checks

To verify the reliability of the estimated results, we carried out robustness checks using three methods. Specifically, we controlled the sample size and the number of explanatory variables and took the family health and family relationship into account. Family health refers to the self-evaluation of family health: 1 for unhealthy and 5 for healthy. Family relationship is a continuous variable measured by the number of meals with family members.

As shown in the second and third columns of Table A1 in Appendix A, children's age (-0.055, $p < 0.01$; -0.098, $p < 0.01$), children's gender (-0.284, $p < 0.01$, for Chinese), family age (0.007, $p < 0.05$; 0.009, $p < 0.01$), family education (0.086, $p < 0.01$; 0.089, $p < 0.01$), education expectation (0.115, $p < 0.01$; 0.157, $p < 0.01$), books (0.105, $p < 0.05$; 0.093, $p < 0.05$), education participation (0.076, $p < 0.01$; 0.054, $p < 0.01$), social communication (0.044, $p < 0.01$; 0.035, $p < 0.01$), medical insurance (-1.450, $p < 0.01$; -1.287, $p < 0.01$), endowment insurance (0.236, $p < 0.01$; 0.188, $p < 0.05$), and tap water (0.092, $p < 0.05$, for Chinese) still have significant influence on children's cognitive ability. Family health (0.038, $p < 0.05$; 0.041, $p < 0.05$) has a positive impact on the two cognitive abilities. Similarly, as shown in the fourth, fifth, sixth, and seventh columns in Table A1 in Appendix A, the significance remains unchanged. Therefore, the results based on 2SLS are robust.

5. Heterogeneity Analysis

The heterogeneity was checked to determine the influence of family factors on children's Chinese and math cognitive abilities.

5.1. Heterogeneity in Gender

As shown in Table A2 in Appendix A, for family culture capital, the influence of family education (0.100, $p < 0.01$; 0.102, $p < 0.01$, for girls) and education participation (0.133, $p < 0.01$; 0.104, $p < 0.01$, for girls) on girls' cognitive ability is greater than that of boys. The influence of family education expectation on girls' (0.157, $p < 0.01$) Chinese cognitive ability is greater than that of boys (0.093, $p < 0.01$), while the influence of family education expectation on boys' (0.162, $p < 0.01$) math cognitive ability is greater than that of girls (0.157, $p < 0.01$). Family books (0.135, $p < 0.1$) only have a significant impact on girls' Chinese cognitive ability. For family social capital, social communication has the greatest impact on girls' cognitive ability (0.054, $p < 0.01$; 0.049, $p < 0.05$, for girls). For social security, medical insurance (-1.958, $p < 0.05$; -1.619, $p < 0.05$, for girls) and endowment insurance (0.298, $p < 0.05$; 0.271, $p < 0.05$, for girls) have the greatest impact on girls' cognitive ability. For living conditions, only tap water has a positive impact on boys' math cognitive ability (0.145, $p < 0.05$). In addition, the larger the family size, the greater the impairment of boys' math cognitive ability. Therefore, the culture capital, social capital, and social security are more sensitive to girls' cognitive ability, while living conditions are more sensitive to boys' cognitive ability.

5.2. Heterogeneity in Urban Location

As shown in Table A3 in Appendix A, for family culture capital, the influence of family education on the cognitive ability of rural children (0.101, $p < 0.01$; 0.116, $p < 0.01$) is greater than that of urban children (0.065, $p < 0.05$; 0.069, $p < 0.05$). Family education expectation has the greatest impact on rural children's math cognitive ability (0.191, $p < 0.01$) and urban children's Chinese cognitive ability (0.123, $p < 0.01$). Family books only affects the math cognitive ability of urban children (0.108, $p < 0.1$). Family education participation has the greatest impact on rural children's Chinese cognitive ability (0.092, $p < 0.01$) and the least impact on urban children's Chinese cognitive ability (0.054, $p < 0.1$). For social communication, the impact on the cognitive ability of rural children (0.057, $p < 0.01$; 0.039, $p < 0.05$) is greater than that of urban (0.041, $p < 0.05$; 0.035, $p < 0.1$). Medical (-1.468, $p < 0.01$; -1.087, $p < 0.05$) and endowment insurance (0.243, $p < 0.05$; 0.193, $p < 0.1$) have a significant impact on the cognitive ability of urban children but not on rural children.

For living conditions, only tap water (0.149, $p < 0.1$) was significant for urban children's Chinese cognitive ability. Therefore, the culture capital and social capital are more sensitive to rural children's cognitive ability, while the social security and living conditions are more sensitive to urban children's cognitive ability.

6. Conclusions

This study used the data from the 2018 China Family Panel Studies to analyze the impact of numerous factors on children's Chinese and math cognitive ability.

Firstly, children's and family's characteristics have significant impact on children's Chinese and math cognitive ability. Among them, children's age, gender, and family size are negative for children's cognitive ability, while family age has a positive impact on children's cognitive ability. Family culture capital, education, education expectation, books, and education participation have a positive impact on children's cognitive ability. For family social capital, the more family social communication, the higher children's cognitive ability. For family living conditions, family use of tap water is more conducive to the improvement of children's cognitive ability. What is more, the influence of family cognitive ability on children's cognitive ability is attenuated by the family capital, which means that the impact of genes are weakened. The above results are based on ordinary least squares (OLS). After introducing instrumental variables Bookiv and Mediv and solving endogeneity, some changes took place in the results. On the one hand, the influence of family books on children's cognitive ability increased significantly. On the other hand, the impact of medical insurance and endowment insurance on children's cognitive ability became significant. Medical insurance was negative, and endowment insurance was positive. In addition, according to the two-stage least squares (2SLS) method, the results are robust after controlling the sample size and increasing the variables.

Moreover, there is heterogeneity in gender and urban location for the influence of numerous factors on children's Chinese and math cognitive ability. In regard to gender, the culture capital, social capital, and social security are more sensitive to girls' cognitive ability, while living conditions are more sensitive to boys' cognitive ability. Specifically, girls' family education, education expectation, books, education participation, social communication, and medical and endowment insurance have a greater impact on cognitive abilities, and tap water is significant for the math cognitive ability of boys. In urban locations, the culture capital and social capital are more sensitive to rural children's cognitive ability, while the social security and living conditions are more sensitive to urban children's cognitive ability. Specifically, rural children's family education, education expectation, education participation, and social communication have a greater impact on cognitive ability, while urban children's family books, medical insurance, endowment insurance, and tap water are more significant for their cognitive ability.

There are some open problems following this research. Due to the imbalance of the initial sample proportion, the proportions of agricultural residence and non-agricultural residence samples were slightly unbalanced after data processing. The heterogeneity in urban location may lead to a slight bias in our full sample model. The error terms of the model may not be independently identically distributed. In addition, there may be further heterogeneity for the influence of numerous factors on children's Chinese and math cognitive ability, and a full mediation analysis should be worthwhile in the future. In this study, we take family cognitive ability as proxy variable of genes, but the empirical results reported in this study are worth checking in full data directly including genetics and environment.

Those findings above provide theoretical support to further narrow the cognitive differences between children.

Author Contributions: Conceptualization, X.D.; methodology, X.D.; analysis, X.D. and W.L.; investigation, X.D. and W.L.; data curation, W.L.; writing—original draft preparation, W.L.; writing—review and editing, X.D.; supervision, X.D.; project administration, X.D.; funding acquisition, X.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Humanities and Social Science Research Fund of the Ministry of Education in China, grant 18YJA790018, in part by the Fundamental Research Funds of the Central Universities, grant CCNU19TS047, in part by the Philosophical and Social Science Research Key Fund of the Department of Education in Hubei Province, grant 17ZD018, and in part by 2018 Graduate Teaching Reform Program at Central China Normal University, grant 2018JG04. The APC was funded by Central China Normal University, China.

Institutional Review Board Statement: Ethical review and approval were waived for this study, due to publically open data from the China Family Panel Studies.

Informed Consent Statement: Patient consent was waived due to publically open data from the China Family Panel Studies.

Data Availability Statement: Data used in this paper can be found from the China Family Panel Studies, http://www.isss.pku.edu.cn/cfps/ (accessed on 13 March 2022).

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

Appendix A

Table A1. Results for robustness tests.

	Robust 1 *N* = **2647**	**Robust 2** *N* = **2647**	**Robust 3** *N* = **2133**	**Robust 4** *N* = **2133**	**Robust 5** *N* = **2133**	**Robust 6** *N* = **2133**
Intercept term	2.869 ***	3.026 ***	2.810 ***	3.020 ***	2.748 ***	2.998 ***
	(0.648)	(0.661)	(0.680)	(0.693)	(0.688)	(0.701)
Child's age	−0.055 ***	−0.098 ***	−0.050 ***	−0.101 ***	−0.050 ***	−0.102 ***
	(0.009)	(0.009)	(0.009)	(0.010)	((0.009)	(0.010)
Child's gender	−0.284 ***	0.003	−0.298 ***	−0.004	−0.299 ***	0.004
	(0.040)	(0.041)	(0.045)	(0.046)	(0.045)	(0.046)
Child's nationality	−0.476	−0.772	−0.408	−0.702	−0.446	−0.750
	(0.461)	(0.470)	(0.457)	(0.466)	(0.459)	(0.468)
Family age	0.007 **	0.009 ***	0.007 **	0.010 ***	0.008 **	0.011 ***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Family gender	−0.036	−0.069	−0.016	−0.045	−0.022	−0.053
	(0.046)	(0.047)	(0.051)	(0.051)	(0.051)	(0.052)
Residence	−0.130 *	−0.119	−0.113	−0.080	−0.116	−0.084
	(0.074)	(0.075)	(0.081)	(0.082)	(0.081)	(0.082)
Urban–rural	−0.056	0.016	−0.070	0.030	−0.065	−0.024
	(0.049)	(0.050)	(0.053)	(0.054)	(0.054)	(0.055)
Family marriage	0.037	0.037	0.085	0.046	0.082	0.047
	(0.101)	(0.103)	(0.108)	(0.110)	(0.108)	(0.110)
Family size	−0.008	−0.011	0.000	−0.002	0.001	−0.001
	(0.011)	(0.011)	(0.012)	(0.012)	(0.012)	(0.013)
Family cognitive ability	−0.021	−0.016	−0.018	−0.017	0.017	−0.016
	(0.016)	(0.016)	(0.017)	(0.018)	(0.018)	(0.018)
Family income	−0.003	0.021	−0.010	0.013	−0.012	0.011
	(0.022)	(0.023)	(0.025)	(0.026)	(0.025)	(0.026)
Children's health investment	−0.002	−0.001	−0.001	−0.006	0.000	−0.004
	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)
Children's education investment	0.014	−0.001	0.014	−0.006	0.014	0.006
	(0.012)	(0.013)	(0.014)	(0.014)	(0.014)	(0.014)
Family education	0.086 ***	0.089 ***	0.080 ***	0.090 ***	0.079 ***	0.088 ***
	(0.017)	(0.017)	(0.018)	(0.018)	(0.018)	(0.019)
Family education expectation	0.115 ***	0.157 ***	0.126 ***	0.174 ***	0.125 ***	0.173 ***
	(0.021)	(0.022)	(0.023)	(0.024)	(0.023)	(0.024)

Table A1. *Cont.*

	Robust 1 N = 2647	Robust 2 N = 2647	Robust 3 N = 2133	Robust 4 N = 2133	Robust 5 N = 2133	Robust 6 N = 2133
Family books	0.105 **	0.093 **	0.094 *	0.084	0.097 *	0.089 *
	(0.046)	(0.046)	(0.051)	(0.052)	(0.051)	(0.052)
Family parenting	0.012	0.075	0.008	0.072	0.004	0.065
	(0.065)	(0.067)	(0.072)	(0.073)	(0.072)	(0.074)
Family education participation	0.076 ***	0.054 ***	0.089 ***	0.066 ***	0.087 ***	0.064 ***
	(0.019)	(0.020)	(0.021)	(0.022)	(0.021)	(0.022)
Family lifestyle	−0.005	−0.028	0.007	−0.032	0.006	−0.034
	(0.026)	(0.027)	(0.029)	(0.030)	(0.029)	(0.030)
Family occupation	−0.013	0.010	−0.015	0.007	−0.014	0.005
	(0.038)	(0.039)	(0.042)	(0.043)	(0.042)	(0.043)
Family information	−0.002	0.020	0.007	0.010	0.006	0.011
	(0.038)	(0.038)	(0.041)	(0.042)	(0.041)	(0.042)
Family human expenditure	−0.006	−0.008	−0.012	−0.010	−0.012	−0.010
	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Family social communication	0.044 ***	0.035 ***	0.046 ***	0.042 ***	0.042 ***	0.036**
	(0.013)	(0.013)	(0.014)	(0.015)	(0.014)	(0.015)
Medical insurance	−1.450 ***	−1.287 ***	−1.447 ***	−1.263 **	−1.474 ***	−1.287 **
	(0.467)	(0.477)	(0.517)	(0.527)	(0.520)	(0.530)
Endowment insurance	0.236 ***	0.188 **	0.233 ***	0.187 **	0.241 ***	0.194 **
	(0.076)	(0.078)	(0.082)	(0.084)	(0.083)	(0.085)
Government support	0.050	0.039	0.074	0.041	0.079	0.047
	(0.045)	(0.046)	(0.049)	(0.050)	(0.049)	(0.050)
Tap water	0.092 *	0.061	0.092 *	0.064	0.092 *	0.064
	(0.048)	(0.049)	(0.053)	(0.054)	(0.053)	(0.055)
Fuel	−0.000	−0.082	0.040	−0.066	0.038	−0.069
	(0.052)	(0.053)	(0.057)	(0.058)	(0.057)	(0.058)
Air purification	−0.076	0.028	−0.157	−0.098	−0.156	−0.099
	(0.113)	(0.116)	(0.128)	(0.130)	(0.128)	(0.131)
Family relationship	0.008	0.001			−0.004	−0.004
	(0.011)	(0.011)			(0.012)	(0.012)
Family health	0.038 **	0.041 **			0.036 *	0.043 **
	(0.018)	(0.018)			(0.020)	(0.020)
R^2	−0.069	0.019	−0.050	0.041	−0.056	0.037
SER	1.017	1.037	1.008	1.027	1.011	1.029

Note: *, **, and *** indicate significance at 10%, 5%, and 1% level, respectively; the standard error is in brackets under the coefficient.

Table A2. Results for two-stage least squares by gender.

	Chinese		Math	
	Boy N = 1429	Girl N = 1218	Boy N = 1429	Girl N = 1218
Intercept term	2.049 *	2.939 ***	1.763	3.517 ***
	(1.158)	(0.941)	(1.185)	(0.949)
Child's age	−0.060 ***	−0.051 ***	−0.097 ***	−0.097 ***
	(0.011)	(0.014)	(0.012)	(0.014)
Child's nationality	0.430	−0.683	0.571	−1.111 **
	(1.013)	(0.550)	(1.037)	(0.554)
Family age	0.005	0.006	0.006 *	0.009 **
	(0.004)	(0.004)	(0.004)	(0.004)
Family gender	−0.049	0.002	−0.103 *	−0.018
	(0.060)	(0.074)	(0.061)	(0.075)
Residence	−0.071	−0.177	−0.066	−0.171
	(0.098)	(0.114)	(0.100)	(0.115)
Urban–rural	−0.049	−0.067	−0.053	0.095
	(0.067)	(0.076)	(0.068)	(0.076)

Table A2. *Cont.*

	Chinese		Math	
	Boy N **= 1429**	**Girl** N **= 1218**	**Boy** N **= 1429**	**Girl** N **= 1218**
Family marriage	−0.108	0.214	−0.049	0.122
	(0.134)	(0.159)	(0.137)	(0.160)
Family size	−0.019	0.007	−0.027 *	0.007
	(0.015)	(0.018)	(0.016)	(0.018)
Family cognitive ability	−0.009	−0.035	−0.003	−0.031
	(0.021)	(0.025)	(0.021)	(0.025)
Family income	−0.002	0.004	0.019	0.035
	(0.031)	(0.034)	(0.031)	(0.034)
Children's health investment	−0.009	0.006	−0.006	0.001
	(0.009)	(0.012)	(0.010)	(0.012)
Children's education investment	0.016	0.009	0.011	−0.020
	(0.017)	(0.021)	(0.017)	(0.021)
Family education	0.077 ***	0.100 ***	0.080 ***	0.102 ***
	(0.022)	(0.027)	(0.022)	(0.027)
Family education expectation	0.093 ***	0.157 ***	0.162 ***	0.157 ***
	(0.027)	(0.038)	(0.027)	(0.039)
Family books	0.076	0.135*	0.088	0.095
	(0.059)	(0.076)	(0.061)	(0.076)
Family parenting	0.110	−0.123	0.233 ***	−0.117
	(0.085)	(0.108)	(0.087)	(0.108)
Family education participation	0.035	0.133 ***	0.015	0.104 ***
	(0.025)	(0.030)	(0.026)	(0.031)
Family lifestyle	−0.012	0.017	−0.038	−0.007
	(0.035)	(0.041)	(0.036)	(0.041)
Family occupation	0.048	−0.053	0.062	−0.028
	(0.052)	(0.058)	(0.053)	(0.058)
Family information	0.039	−0.066	0.056	−0.025
	(0.051)	(0.062)	(0.052)	(0.063)
Family human expenditure	−0.010	−0.006	−0.015	−0.002
	(0.014)	(0.016)	(0.014)	(0.016)
Family social communication	0.040 **	0.054 ***	0.035 **	0.049 **
	(0.017)	(0.021)	(0.017)	(0.021)
Medical insurance	−1.124 **	−1.958 **	−1.151 **	−1.619 **
	(0.560)	(0.819)	(0.573)	(0.825)
Endowment insurance	0.186 **	0.298 **	0.127	0.271 **
	(0.090)	(0.134)	(0.092)	(0.135)
Government support	0.021	0.089	0.060	0.004
	(0.057)	(0.076)	(0.058)	(0.077)
Tap water	0.074	0.098	0.145 **	−0.042
	(0.064)	(0.077)	(0.065)	(0.077)
Fuel	−0.003	−0.002	−0.074	−0.090
	(0.068)	(0.082)	(0.070)	(0.083)
Air purification	0.082	−0.271	0.108	−0.080
	(0.155)	(0.174)	(0.158)	(0.175)
R^2	−0.011	−0.227	0.072	−0.058
SER	0.987	1.078	1.010	1.086

Note: *, **, and *** indicate significance at 10%, 5%, and 1% level, respectively; the standard error is in brackets under the coefficient.

Table A3. Results for two-stage least squares by urban location.

	Chinese		Math	
	Urban N **= 1141**	**Rural** N **= 1506**	**Urban** N **= 1141**	**Rural** N **= 1506**
Intercept term	2.069 **	3.458 ***	1.815 *	4.385 ***
	(0.897)	(1.123)	(0.940)	(1.179)
Child's age	−0.059 ***	−0.054 ***	−0.087 ***	−0.104 ***
	(0.013)	(0.012)	(0.013)	(0.012)
Child's gender	−0.256 ***	−0.318 ***	−0.018	0.011
	(0.059)	(0.054)	(0.062)	(0.057)
Child's nationality	−0.404	−0.445	−0.300	−1.014
	(0.716)	(0.606)	(0.750)	(0.637)
Family age	0.008 *	0.004	0.012**	0.006
	(0.005)	(0.003)	(0.005)	(0.004)
Family gender	−0.086	0.009	−0.044	−0.065
	(0.068)	(0.063)	(0.071)	(0.066)
Residence	−0.171 *	−0.015	−0.179 *	0.049
	(0.089)	(0.163)	(0.093)	(0.171)
Family marriage	0.047	0.079	0.068	0.019
	(0.151)	(0.137)	(0.158)	(0.144)
Family size	−0.010	−0.011	−0.004	−0.019
	(0.017)	(0.014)	(0.018)	(0.015)
Family cognitive ability	−0.005	−0.029	0.004	−0.031
	(0.022)	(0.022)	(0.023)	(0.023)
Family income	0.050	−0.040	0.059 *	−0.002
	(0.032)	(0.031)	(0.034)	(0.033)
Children's health investment	−0.006	−0.001	0.010	−0.011
	(0.011)	(0.010)	(0.011)	(0.011)
Children's education investment	0.002	0.023	0.002	−0.002
	(0.019)	(0.016)	(0.020)	(0.017)
Family education	0.065 **	0.101 ***	0.069 **	0.116 ***
	(0.027)	(0.025)	(0.029)	(0.026)
Family education expectation	0.123 ***	0.112 ***	0.110 ***	0.191 ***
	(0.034)	(0.028)	(0.035)	(0.030)
Family books	0.087	0.093	0.108 *	0.054
	(0.062)	(0.069)	(0.065)	(0.072)
Family parenting	0.100	−0.026	0.120	0.055
	(0.101)	(0.086)	(0.106)	(0.090)
Family education participation	0.054 *	0.092 ***	0.055 *	0.058 **
	(0.031)	(0.025)	(0.032)	(0.026)
Family lifestyle	0.036	−0.028	−0.032	−0.010
	(0.038)	(0.037)	(0.039)	(0.039)
Family occupation	0.061	−0.097 *	0.064	−0.067
	(0.051)	(0.056)	(0.053)	(0.059)
Family information	0.033	−0.020	0.042	0.006
	(0.055)	(0.052)	(0.058)	(0.055)
Family human expenditure	−0.008	−0.005	−0.010	−0.009
	(0.015)	(0.014)	(0.016)	(0.015)
Family social communication	0.041 **	0.057 ***	0.035 *	0.039 **
	(0.020)	(0.017)	(0.021)	(0.018)
Medical insurance	−1.468 ***	−1.300	−1.087 **	−1.861
	(0.478)	(1.126)	(0.501)	(1.181)
Endowment insurance	0.243 **	0.200	0.193 *	0.198
	(0.096)	(0.140)	(0.100)	(0.147)
Government support	0.070	0.040	0.105	0.016
	(0.067)	(0.067)	(0.070)	(0.071)
Tap water	0.149 *	0.069	0.094	0.083
	(0.084)	(0.064)	(0.089)	(0.067)
Fuel	0.097	−0.004	−0.060	−0.075
	(0.100)	(0.067)	(0.105)	(0.071)
Air purification	−0.110	0.094	0.018	0.072
	(0.131)	(0.220)	(0.138)	(0.231)
R^2	−0.059	−0.035	0.018	−0.043
SER	0.978	1.030	1.025	1.081

Note: *, **, and *** indicate significance at 10%, 5%, and 1% level, respectively; the standard error is in brackets under the coefficient.

References

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association* 91: 444–55. [CrossRef]

Autor, David H. 2014. Skills, education, and the rise of earnings inequality among the 'other 99 percent'. *Science* 344: 843–51. [CrossRef] [PubMed]

Bourdieu, Pierre, and Loic J. D. Wacquant. 1992. An Invitation to Reflexive Sociology. *Contemporary Sociology* 22: 450.

Bourdieu, Pierre, and Jean Clacude Passeron. 1977. Reproduction in Education, Society and Culture. *British Journal of Sociology* 30: 237–41.

Brookhart, M Alan, Philip S. Wang, Daniel H. Solomon, and Sebastian Schneeweiss. 2006. Instrumental variable analysis of secondary pharmacoepidemiologic data. *Epidemiology* 17: 373–74. [CrossRef] [PubMed]

Cao, Dandan, Shengquan Luo, Xiaoping Yang, and Wentao Wang. 2018. Effect of internet on the development of cognitive ability of urban and rural teenagers. *China Educational Technology* 2018: 9–17.

Chen, Xi, Xiaobo Zhang, and Xin Zhang. 2017a. Smog in Our Brains: Gender Differences in the Impact of Exposure to Air Pollution on Cognitive Performance. *GLO Discussion Paper Series* 2017: 115–57. [CrossRef]

Chen, Yvonne, Li Li, and Xiao Yun. 2017b. Early Life Exposure to Tap Water and the Development of Cognitive Skills. *Social Science Electronic Publishing* 17: 49. [CrossRef]

Chen, Hua, Jing Zhao, and Yujia Shao. 2020. Will basic medical insurance system change parents' educational expectations of their children?A study based on CFPS data of rural household. *Financial Economics Research* 35: 143–58.

Chen, Haiyan, Li Chen, and Guang Hao. 2021. Sex difference in the association between solid fuel use and cognitive function in rural China. *Environmental Research* 195: 110820. [CrossRef]

Cong, Xiaowei, Juan Zhang, Rongli Sun, and Y. Pu. 2021. Indoor unclean fuel cessation linked with adult cognitive performance in China. *Science of The Total Environment* 775: 145518. [CrossRef] [PubMed]

Cui, Peng, and Athey Susan. 2022. Stable learning establishes some common ground between casual interference and machine learning. *Nature Machine Intelligence* 4: 110–15. [CrossRef]

Dong, Fang, and Jiangtao Zhou. 2019. Study on the effect and heterogeneity of parental time investment on their children's human capital. *Northwest Population Journal* 40: 48–61.

Fan, Jingbo, Lingdong Meng, and Xiaoping Yang. 2019. A comparative study on the trend differences of family factors affecting children's higher education enrollment opportunities (in Chinese). *China Higher Education* 2019: 58–60.

Fang, Changchun, and Xiaotian Feng. 2005. How distinction of social stratum affects the attainment of education: An analysis on split-flows of education. *Tsinghua Journal of Education* 26: 22–30.

Fang, Guangbao, and Yi Hou. 2019. How family social economic status impact the development of secondary students' cognitive competence. *Global Education* 48: 68–76.

Fang, Chao, and Bin Huang. 2020. Matthew effect or equity effect: Heterogeneity test of family education expenditure and inequality of educational outcome. *Education & Economy* 36: 58–67.

Fang, Kuangnan, and Ziyi Zhang. 2013. Study on effects on social security on household consumption. *Statistical Research* 30: 51–58.

Fang, Chao, Diyang Zeng, and Bin Huang. 2020. Family size, sibling structure and educational attainment of school-age children: Evidence from the survey of CEPS. *Journal of Central China Normal University (Humanities and Social Sciences)* 59: 181–92.

Farkas, George. 2003. Cognitive Sills and Noncognitive Traits and Behaviors in Stratification Processes. *Annual Review of Sociology* 29: 541–62. [CrossRef]

Gao, Yaqing, Long Zhang, Ashish Kc, Yinping Wang, Siyu Zou, Chunyi Chen, Yue Huang, Xiaoyi Mi, and Hong Zhou. 2021. Housing environment and early childhood development in sub-Saharan Africa: A cross-sectional analysis. *PLoS Medicine* 18: e1003578. [CrossRef] [PubMed]

Geng, Zhi. 2004. Observational studies and confounding factors (in Chinese). *Journal of Statistics and Information* 2004: 13–17.

Granovetter, Mark S. 1973. The Strength of Weak Ties. *American Journal of Sociology* 78: 1360–80. [CrossRef]

Gu, Hongwei, and Qiuping Yang. 2013. Income, expectation and education expenditure: An empirical analysis of the current Chinese family investment in Education (in Chinese). *Macroeconomics* 2013: 68–74+88.

Guo, Congbin, and Weifang Min. 2006. The effect of familial economical and cultural capital on educational attainment in China. *Journal of Higher Education* 27: 24–31.

Guo, Jinguang, and Hao Sun. 2019. Will social security help poverty reduction in the future? *Study and Practice* 2019: 105–17.

Hao, Juan. 2018. A comparative research on difference and its trend of education between genders and between urban and rural areas. *Education Science* 34: 20–25.

Hong, Zhichao, and Miao Zhang. 2021. The impact of family capital on education quality and its mechanism—An Empirical Study Based on CEPS survey data (in Chinese). *Finance and Economy* 2021: 90–96.

Hong, Yanbi, and Yandong Zhao. 2014. From capital to habitus: The class differentiation of family educational pattern in urban China. *Sociological Studies* 29: 73–93+243. [CrossRef]

Hou, Wenpeng, Tongxing Tan, and Yujie Wen. 2020. The effect of increased family finance and dual-parental absence since infancy on Children's cognitive Abilities. *Social Science & Medicine* 266: 113361.

Houmark, Mikkel Aagaard, Victor Ronda, and Michael Rosholm. 2020. The Nurture of Nature and the Nature of Nurture: How Genes and Investments Interact in the Formation of Skills. *IZA Discussion Papers* 2020: 13780. [CrossRef]

Hu, Xiaoli, and Dandan Xie. 2011. Research on problem-oriented strategies to promote online learners' cognitive ability skills. *Journal of Distance Education* 29: 21–26.

Huang, Chao. 2018. Parenting styles and the development of non-cognitive skills among Chinese adolescents. *Chinese Journal of Sociology* 38: 216–40.

James, Richard. 2000. Non-traditional Students in Australian Higher Education:Persistent Inequities and the New ldeology of Student Choice. *Tertiary Education and Management* 6: 105–18. [CrossRef]

Jiang, Jiajiang, and Fan Zhang. 2020. Semi-disembedding growth: Family structure and gender differences in adolecent development. *Zhejiang Academic Journal* 2020: 142–53.

Jin, Jiuren. 2019. From capital gap to field separation: A study of the family support gap in urban and rural education. *Education Science* 35: 1–8.

Kleinjans, Kristin J. 2010. Family Background and Gender Differences in Education Expectations. *Economic Letters* 107: 125–27. [CrossRef]

Kuang, Xiaofang, Weiwei Yan, Yuting Chen, Qinge Wang, and Juan Yang. 2019. Research on Chinese primary-school students' second language Learning. *Modern Educational Technology* 29: 72–79.

Li, Yanan. 2012. The effects of family income on children education:An empirical analysis of CHNS data. *South China Population* 27: 46–53+45.

Li, Jiali. 2017. Effects of parental involvement and intergeneration closure on student cognitive ability: Focusing on Coleman's social capital. *Research in Educational Development* 37: 6–14.

Li, Jiaoyuan, and Xiangming Fang. 2019. Impact and mechanism of parental health on children's academic performance. *Journal of Harbin Institute of Technology (Social Sciences Edition)* 21: 54–62.

Li, Jiaoyuan, and Zheng Shen. 2021. Parental mental health and children's human capital accumulation in rural China: Evidence from China Family Panel Studies (CFPS). *Northwest Population Journal* 42: 71–84.

Li, Yaxian, and Chuanchuan Zhang. 2018. Cognitive ability and consumption: A new perspective of understanding the high saving rate of elderly people. *Economics Perspectives* 2018: 65–75.

Li, Li, and Wenlong Zhao. 2017. The influence of family background and cultural capital on cognitive ability and non cognitive ability. *Dongyue Tribune* 38: 142–50.

Liang, Chen. 2020. On the influence of family capital on children's educational attainment (in Chinese). *China Collective Economy* 2020: 161–62.

Liang, Wenyan, Xiaomei Ye, and Tao Li. 2018. How does parental involvement affect the cognitive ability of migrant children: An empirical study based on CEPS database. *Journal of Educational Studies* 14: 80–94.

Lin, Nan. 2005. *Social Capital: A Theory of Social Structure and Action*. Shanghai: Shanghai People's Publishing House.

Lin, Xin, Jingyu Xie, and Suxu Lin. 2021. The influence of family capital on rural children's academic achievements—An empirical research based on the data of CFPS(2018). *Theory and Practice of Education* 41: 24–30.

Liu, Shenglong, and Tianyu Jin. 2020. Does birth quantity affect children's educational attainment? Evidence from Chinese population census. *The Journal of World Economy* 43: 121–43.

Liu, Dedi, and Zengxin Xue. 2021. The educational spillover effect of the minimum living guarantee system for rural residents: Empirical analysis based on the human capital for poor children. *Northwest Population Journal* 42: 44–56.

Liu, Baozhong, Yueyun Zhang, and Jianxin Li. 2015. Family SES and adolescent educational expectation: Mediating role of parental involvement. *Peking University Education Review* 13: 158–76.

Miettinen, Olli S., and E. Francis Cook. 1981. Confounding: Essence and detection. *American Journal of Epidemiology* 114: 593–603. [CrossRef] [PubMed]

Nauze, Andrea La, and Edson R. Severnini. 2021. Air Pollution and Adult Cognition: Evidence from Brain Training. *NBER Working Papers* 2021: 14353.

Plomin, Robert, and Sophie von Stumm. 2018. The New Genetics of Intelligence. *Nature Reviews Genetics* 19: 148. [CrossRef] [PubMed]

Putnam, Robert D. 2000. *Bowling Alone—The Decline and Revival of American Communities*. Beijing: Peking University Press, vol. 2011, pp. 343–55.

Ronda, Victor, Esben Agerbo, Dorthe Bleses, Preben Bo Mortensen, Anders Børglum, David M. Hougaard, Ole Mors, Merete Nordentoft, Thomas Werge, and Michael Rosholm. 2020. Family Disadvantage, Gender and the Returns to Genetic Human Capital. *IZA Discussion Papers* 2020: 13441. [CrossRef]

Saasa, Sherinah K. 2018. Education among Zambian children: Linking head of household characteristics to school attendance. *Vulnerable Children and Youth Studies* 13: 239–46. [CrossRef]

Schikowski, Tamara, and Hicran Altug. 2020. The role of air pollution in cognitive impairment and decline. *Neurochemistry International* 136: 104708. [CrossRef] [PubMed]

Shen, Ji. 2019. The impact of health on children's cognitive ability. *Youth Studies* 2019: 14–26+94.

Silventoinen, Karri, Aline Jelenkovic, Reijo Sund, Antti Latvala, Chika Honda, Fujio Inui, Rie Tomizawa, Mikio Watanabe, Norio Sakai, Esther Rebato, and et al. 2020. Genetic and environmental variation in educational attainment: An individual-based analysis of 28 twin cohorts. *Scientific Reports* 10: 12681. [CrossRef]

Tao, Dongjie. 2019. Siblings size and youth's cognitive ability: Resource dilution or parental selection? *Education & Economy* 35: 29–39.

Teacherman, Jay D. 2000. Parental Cultural Capital and Educational Attainment in the Nether Lands a Refinement of the Cultural. *American Sociological Review* 73: 92–11.
Wang, Chuanyan, and Zuwang Chu. 2019. The impact of sense of family belonging on migrant children's academic performance: The mediating effect of parent-child conflicts. *Chinese Journal of Special Education* 2019: 61–68.
Wang, Pengcheng, and Xin Gong. 2020. The influence of household cultural capitals on preschool attendance: An empirical research based on CFPS survey data. *Studies in Early Childhood Education* 2020: 43–54.
Wang, Chunchao, and Junjie Lin. 2021. Parental companionship and children's human capital development. *Educational Research* 42: 104–28.
Wang, Fuqin, and Yiwen Shi. 2014. Family background, educational expectation and college degree attainment: An empirical study based on Shanghai survey. *Chinese Journal of Sociology* 34: 175–95.
Wei, Wei, Yifang Wu, Ping Ren, and Liang Luo. 2015. Predictors of parental involvement: Family social economic status and parents' psychological factors. *Journal of Beijing Normal University (Social Sciences)* 247: 62–70.
Wu, Jia, Jiada Lin, and Xiao Han. 2020. Parental patience, parental style and children's human capital accumulation. *Economics Perspectives* 2020: 37–53.
Wu, Jia, Guansheng Wu, and Biao Li. 2021. Can early-life health input boost children's long-run cognitive ability? *China Economic Quarterly (in Chinese)* 21: 157–80.
Xie, Yuxiang, and E. Xie. 2019. Educational, occupational mobility and intergenerational socioeconomic status transmission. *Chinese Journal of Population Science* 2019: 40–52+126–127.
Xue, Haiping. 2018. Family capital and education attainment: Analysis on the mediating effect of shadow education. *Education & Economy* 2018: 69–78.
Xue, Xiaoyuan, and Rongxiang Cao. 2004. Cultural capital, cultural products and cultural system—Cultural capital theory after Bourdieu (In Chinese). *Marxism & Reality* 2004: 43–49.
Xue, Zengxin, Zhipeng He, Zhengshun Qi, and Dedi Liu. 2021. The influence of new rural pension scheme on the cognitive ability of rural left-behind children. *World Agriculture* 2021: 83–93.
Yan, Bohan. 2017. Effects of rural-to-urban migration to children's cognitive ability development in China: Analysis based on census data of urbanization and migration in 2012. *Chinese Journal of Sociology* 37: 59–89.
Yang, Baoyan, and Minggang Wan. 2015. How father's education and economic capital influence academic achievement: Analysis of mediation and moderation effects. *Peking University Education Review* 13: 127–145+192.
Yang, Rudai, and Bishu Yuan. 2019. New rural pension scheme and consumption of rural residents. *Consumer Economics* 35: 3–12.
Yang, Hong, and Ke Zhang. 2020. Cognitive abilities, social interaction way and household portfolio choice—Empirical analysis based on CFPS data. *Review of Investment Studies* 39: 67–81.
Yao, Hao, and Zhong Ye. 2018. Family background, education quality and development of student ability: Multilayer linear model analysis based on CEPS. *Contemporary Education and Culture* 10: 70–79.
Yin, Jing, and Lin Fan. 2021. Early childhood language cognition, language development and socioeconomic status. *Technology Enhanced Foreign Language Education* 2021: 109–116+16.
Zhang, Chunni. 2020. The long-term influence of marital breakdown in divorced families upon their children's socioeconomic achievements in China. *Journal of Peking University (Philosophy and Social Sciences)* 57: 128–39.
Zhang, Yuehua, and Tong Li. 2021. The impact of cognitive ability on participation in the new rural pension program—an empirical study based on China Family Panel Studies. *Insurance Studies* 2021: 89–98.
Zhang, Wenhong, and Di Su. 2018. Cultural capital, economic capital and stratum reproduction. *Jianghai Academic Journal* 2018: 102–12.
Zhang, Xiyang, Lu Leng, Honhjun Chen, Xiaoyi Fang, Zeng Shu, and Xiuyun Lin. 2017. Parental rearing pattern mediates the association between social economic status and cognitive ability of migrant children. *Psychological Development and Education* 33: 153–62.
Zhao, Liange, Xinjie Deng, and Xueyuan Wang. 2018. Socioeconomic status, environmental sanitation facilities and health of rural residents. *Issues in Agricultural Economy* 2018: 96–107.
Zheng, Lei, Xiang Qi, and Yuna Hou. 2018. The intergenerational effects of family on education: Theories, methodologies and evidences. *Journal of Social Development* 5: 177–202+245–246.
Zheng, Lei, Xiang Qi, Zhiyong Zhu, and Dingquan Zhang. 2021. Home internet access and urban-rural cognition gap of middle school students. *Research in Educational Development* 41: 10–18.
Zhou, Jinyan, and Xue Zou. 2016. Comparing the private tutoring options between students in China and the United State—Evidence form 2012 PISA survey and investigation. *Education & Economy* 2: 44–52.
Zhou, Mi, Xiaotong Sun, Zhuang Kang, and Li Huang. 2019. Influence of parents' educational expectation expectancy on rural children's cognitive ability: An empirical study based on the CFPS Panel data. *Journal of Hunan Agricultural University (Social Sciences)* 20: 57–62+90.
Zhou, Chunfang, Qun Su, and Xue Chang. 2021. Study on shadow education participation of rural migrant children and its effect on education equalization. *Journal of Agrotechnical Economics* 2021: 130–144.
Zhu, Hong, and Wenjie Zhang. 2020. Elite college students' family backgrounds and their development—A survey on freshmen of Peking University (2016–2018). *Journal of Higher Education* 41: 71–82.

Zhu, Jian, Lei Xu, and Hui Wang. 2018. Research on urban-rural differences in education transition between generation-Evidence from CGSS data. *Education &Economy* 2018: 45–55.

Zimmer, Zachary, Linda G. Martin, Mary Beth Ofstedal, and Yi-Li Chuang. 2007. Education of Adult Children and Mortality of Their Elderly Parents in Taiwan. *Demography* 44: 289–304. [CrossRef]

Journal of
Intelligence

MDPI

Article

Playful Testing of Executive Functions with Yellow-Red: Tablet-Based Battery for Children between 6 and 11

Ricardo Rosas [1,2,*], Victoria Espinoza [1,2], Camila Martínez [2] and Catalina Santa-Cruz [2]

[1] Centro de Desarrollo de Tecnologías de Inclusión, Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago 7820436, Chile
[2] Centro de Justicia Educacional, Facultad de Educación, Pontificia Universidad Católica de Chile, Santiago 7820436, Chile
* Correspondence: rrosas@uc.cl

Abstract: Executive functions are psychological processes of great importance for proper functioning in various areas of human development, including academic performance. For this reason, from both clinical and educational perspectives, there is great interest in how they are assessed. This article describes the development and standardization process of Yellow-Red, an instrument for directly assessing executive functions in children between 6 and 11 years of age in a playful format using digital support. The test was based on a three-factor model of executive functioning: inhibition, working memory, and cognitive flexibility. Yellow-Red comprises six subtests: cognitive inhibition, behavioral inhibition, auditory working memory, visual working memory, cognitive flexibility, and a global assessment test of executive functions. The test was administered to 245 boys and girls between 6 and 11 years of age. Along with the Yellow-Red subtests, gold standard tests were applied for each of the executive functions assessed. The test's psychometric properties are powerful in both reliability and validity evidence. The reliability indices are all greater than 0.8. As evidence of convergent validity, correlations were established between the tests, and the tests considered gold standards. All correlations were significant, with values ranging between 0.42 and 0.73. On the other hand, the factor structure of the test was analyzed using confirmatory factor analysis. Although it is possible to demonstrate the progressive differentiation of the factor structure with age, it was only possible to find two factors at older ages, one for inhibition/flexibility and one for working memory.

Keywords: executive functions; technology-based assessment; cognitive assessment

Citation: Rosas, Ricardo, Victoria Espinoza, Camila Martínez, and Catalina Santa-Cruz. 2022. Playful Testing of Executive Functions with Yellow-Red: Tablet-Based Battery for Children between 6 and 11. *Journal of Intelligence* 10: 125. https://doi.org/10.3390/jintelligence10040125

Received: 14 October 2022
Accepted: 5 December 2022
Published: 14 December 2022

1. Introduction

Executive functions are the cognitive abilities responsible for planning, controlling, and guiding thoughts, feelings, and actions. They are the central executive of the cognitive system, i.e., they transform intentions and purposes into practical actions. People with a greater development of executive functions are more likely to achieve their goals, as they can plan their tasks adequately. There is ample evidence of the impact of executive functions on various areas of human development, especially academic performance (Diamond 2016). Next, we define the theoretical model on which the Yellow-Red test was built, describe the relationship between children and technology use, and summarize previous contributions regarding the assessment of executive functions at the international level. Subsequently, a general description of the test is presented, allowing a better understanding of the results and reflections derived from the standardization process of the Yellow-Red test.

1.1. The Three-Component Model of Executive Functions

There are various models for conceptualizing executive functions; one of the most widely accepted is the one that defines the presence of three basic components of executive functions. These basic components develop interdependently during childhood and early

adolescence and serve as the basis for developing higher-order executive functions such as planning and problem-solving.

Inhibitory control includes inhibiting thoughts, actions, or behaviors in the face of competing for internal or external stimuli. It is, therefore, an ability that allows the inhibition of cognitive, emotional, and behavioral factors. The first and second are related to thinking and memory. The third, with behavioral inhibition, for example, is related to gratification delay.

Working memory is the ability to operate with mental representations, whether visual, auditory, or episodic. According to Cowan (2017), it is a set of components of the mind that hold a limited amount of information temporarily available for processing the task at hand. It is an ability of limited capacity, although it progressively expands with age, reaching its maximum capacity around age 12.

Finally, cognitive flexibility is the ability to provide alternative solutions to the same problem. It is closely related to creativity, and of the three functions, it is the one with the latest appearance (Diamond 2013).

The three main components of EF, as shown in Figure 1, can be understood as successively integrated over time. The first to make its appearance is inhibitory control, followed by working memory, and finally, cognitive flexibility. Although, conceptually, the three components are present throughout development, it is clear that their nature changes from 18 months of age, which is the age at which language begins to be a fundamental part of cognitive development. Language plays a fundamental role in executive functioning since it allows for labeling internal instructions that inhibit behaviors and actions, processing problems, and seeking alternative solutions to unknown problems.

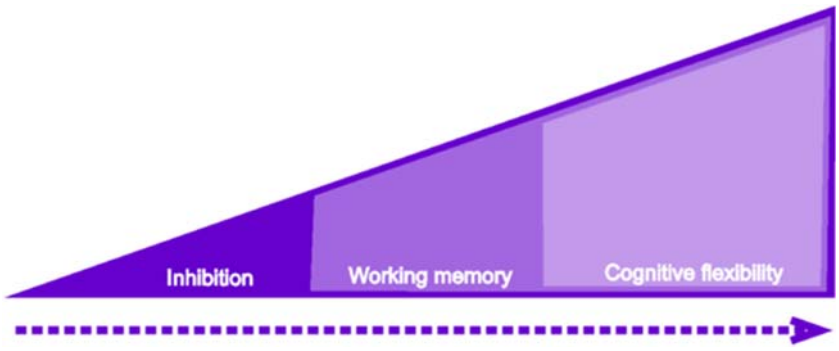

Figure 1. The components that together make up executive functions are shown.

The three components of EF also have important interrelationships, given by their successive integration into development. For example, to solve problems in working memory, it is essential to have active interference control of internal and external stimuli while processing the solution, which is a component of inhibitory control. For this reason, many models of working memory (e.g., Kane and Engle 2003) incorporate interference control as a component of working memory. However, it could properly be considered as a factor of inhibitory control. As Diamond (2013) noted, some authors (e.g., Baddeley and Hitch 1994) incorporated inhibitory and flexibility factors in their working memory model. However, following this author, we kept the three factors separate in the present work, as Miyake et al. (2000) suggested.

Likewise, cognitive flexibility requires both working memory and inhibitory control to provide the alternative solution being processed (e.g., if the task is to say all the words that begin with a given letter, in working memory, I must simultaneously evaluate that the new term I come up with actually starts with that letter and simultaneously remember and discard repeating the ones I have already said).

An interesting issue regarding the progressive differentiation of executive functions is the unicity versus diversity approach formulated by Friedman and Miyake (2017). Essentially, this approach postulates that executive functions show unicity and diversity,

depending on both the analysis techniques used and the developmental level of the samples assessed. We stayed with the latter aspect for the present article, which focuses mainly on children. Friedman and Miyake (2017) reported that, although some studies show a unicity of executive functions at early ages, all studies evidence a differentiation of working memory and flexibility in older children or adults. In other words, at earlier ages, the appropriate mode for understanding executive functions is unity, while at older ages, it is that of diversity. The evidence on this point is mixed, specifically regarding when and which components are part of the first and second factors in school-aged children. This is partly due to the variety of tests used, as the selection of tests according to each component of executive functions is highly heterogeneous, both in their assessment objectives and the way they are assessed (for a comprehensive review on this topic, see Lee et al. 2013).

1.2. Use of Technology Tools for Child Assessment

New generations grow up immersed in digital media-rich environments, and technology is integral to their lives from birth (Sweeney and Geer 2008). From a very young age, children are exposed to technological resources, which translates into an early mastery of various digital tools (Mcmanis and Gunnewig 2012).

The use of technological tools has increased both in the world and in Chile. In the USA, in 2018, 85% of households had an internet connection (United States Census Bureau 2021). In Chile, the statistics are similar, with 87.4% of households reported to have an internet connection in 2017. Likewise, access to technological devices is equally high, with 85.7% of households in which school-age children live have a smart mobile device (Subsecretaría de Telecomunicaciones de Chile 2017). According to Chaudron et al. (2018), who investigated the use of technology in children aged 0–8 years in 21 countries, the use of digital technologies starts earlier and earlier (under two years), and tablets and smartphones are the preferred devices of children, due to their multifunctionality and portability. According to these authors, devices with touch screens are appreciated by children, especially for their ease of use, the possibility of accessing different applications, and their playful aspects.

For these reasons, the need to incorporate technology into educational systems has been raised (Sweeney and Geer 2008), considering its use both for the mediation of teaching and for the assessment of learning. Day et al. (2019) noted that there is a need for technology- and game-based executive function assessment tools that can be used outside of the clinical or academic context, allowing for accurate, ecological, and contingent assessments.

Technology-mediated assessments have several advantages over traditional assessments, as they allow for gamification of the assessment format by incorporating aspects traditionally related to video games or applications. They also allow the standardization of certain technical elements, such as instructions, examples, or forms of response, and the automation of correction processes. On the other hand, it was observed that the use of technological instruments allows the assessment of aspects impossible to assess and apply in pencil and paper instruments, for example, reaction times, presentation of algorithmically programmed items, and the measurement of aspects related to behavior (Germine et al. 2019; Parsey and Schmitter-Edgecombe 2013).

Parsey and Schmitter-Edgecombe (2013) noted that the widespread access of younger generations has produced a cohort effect, in which children and young people perform better on computer-based assessments than older people with less technological experience. Moreover, the use of technology in assessment contexts generates an increased level of student engagement and motivation, enabling the expression of their full performance potential (Perrotta et al. 2013; Rosas et al. 2015).

Germine et al. (2019) recommended focusing on four aspects when developing technology-based assessment instruments: (a) designing interfaces that are accessible and appealing to the target age group, (b) developing simple and clear instructions, (c) presenting applied test items that allow the user to interact with the test, which is more effective

than reading written instructions, and (d) developing specific norms for technology-based tests rather than adapting norms from pencil-and-paper instruments.

However, it is important to consider whether technology-mediated assessments correspond to those in traditional formats. In this regard, several meta-analyses involving tests with students from K to 12 have shown no significant differences in the results of the two types of assessment (Kingston 2009; Wang et al. 2008), which contributes to the reliability of this type of instrument.

1.3. Description of Instruments and Gold Standards for FE Assessment

Multiple research fields have approached executive functions, such as neuropsychology, cognitive psychology, education, and, more recently, cognitive neurosciences. Likewise, each of these areas has developed its assessment paradigms, depending on the nature of their studies.

The first research came from neuropsychology and was based mainly on studying adults with some type of brain injury, thus establishing the relationship between executive functions and the frontal lobe. The works of Luria and his collaborators were paradoxical. They described frontal lobe syndrome in 1964, proposing a series of tasks to evaluate the relationship between neurological disorders and performance in cognitive and motor functions (Canavan et al. 1985). Thus, in 1980, the standardized version of their procedures was published as the so-called Luria-Nebraska battery. In 1981, they presented the first version for children between 8 and 11 years of age (Plaisted et al. 1983). According to Zelazo et al. (2016), interest in assessing executive functions in children only arises when the belief that the limited frontal lobe development during childhood was demystified in the early 1980s. It has been shown that it is just the opposite since it has been demonstrated that the frontal lobe shows a more significant development during childhood.

From this new interest in the association between executive functions and frontal lobe development in children, children's versions of instruments used to assess executive functions in adults were developed. An example is the Stroop Interference Test (Stroop 1935), one of the most widely used neuropsychological measures. The Stroop test consists of three consecutive tasks: First, a list of colors expressed in words must be read aloud. Second, one must name a series of colors presented as such in rectangles. Third, a list of colors printed in ink of a color different from that expressed by the word, for example, "yellow" printed in red ink, must be named. According to Homack and Riccio (2004), there is consensus that this last task measures cognitive flexibility and inhibition. The original version proposed by Stroop was interpreted in different ways. One of the most widely used versions is the one developed by Golden (1978), standardized in 2003 for adults and 2002 for children. The children's version can be applied to children aged from five to fourteen years and only differs from the adult version in scoring norms (Moran and Yeates 2018; Rozenblatt 2018). This is an example of how the original tools were designed for application with adults. Their children's versions are only later adaptations, not instruments directly created for these age groups, which do not have standardized versions. If they do, they present very poor application norms (Carlson 2005). According to Hughes and Ensor (2011), child adaptation from adult tasks runs the risk of losing critical components of executive functions, for example, oversimplifying them or not considering other cognitive aspects that develop in parallel or later, such as the use of language, specifically vocabulary.

Since 2000, research in psychology and neuroscience has grown exponentially, generating several instruments consisting of individual behavioral tasks based on performance (for more details, see Carlson (2005) and Garon et al. (2008)), which have become more accurate thanks to their technological versions applied on PCs and, later, on tablets. One of the most widely used tests is the Hearts and Flowers test, which corresponds to a version of "Dots" originally developed by Davidson et al. (2006). This test consists of three consecutive tasks; in the first block, the person must press a key on the same side on which a heart appears (congruent block); in the second task, he/she must press a key on the opposite side of which a flower appears (incongruent block). Finally, there is a mixed block in which

hearts and flowers appear randomly. The individual must follow two rules simultaneously, depending on the stimulus that appears, forming a mixed block. Despite its wide use, Hearts and Flowers does not have norms, validity, or reliability studies (Camerota et al. 2020).

In 1996, Zelazo et al. (1996) presented the first version of the Dimensional Change Card Sort (DCCS). in which the child is asked to sort a series of drawings, first according to their shape (put a card with a rabbit on top of another rabbit card, regardless of its color) and then according to their color (put a card with a red figure on top of another red card, regardless of its shape). The child must sort 48 cards according to the instruction of the evaluator, who randomly says "shape" or "color." The DCCS is now part of a free, validated, norm-referenced battery for the North American population aged 2.5 to 85 (Zelazo et al. 2013). This is a digital version, whose only disadvantage is that it is only available for IOS devices. The previous tests are traditionally laboratory-based but more ecological; behavioral measures are generally related to cognitive and educational psychology. It is in these contexts where tests that have not been standardized but are widely used are also used, such as Simon says (Strommen 1973), based on the traditional children's game, or Head Shoulders Knees and Toes (Cameron Ponitz et al. 2008), in which the child is progressively asked to touch parts of his body in alternating order.

Other instruments that can be used for the assessment of executive functions are the ENFEN (Portellano et al. 2011), which assesses the global maturational development of children between 6 and 12 years of age with the main focus on executive functions. This test presents norms for the Spanish population with an individual application format with attractive tasks for students, does not directly consider a play format, and does not use digital support. On the other hand, an alternative is the Psychology Experiment Building Language (PEBL) platform, which allows the free programming of digital tests. This platform has some traditional tests pre-designed on the platform, focusing on evaluating executive functions. Among the tests that can be selected is a version of Berg's Card Sorting Test, similar to the Wisconsin test, Corsi's block test, and an implementation of Eriksen's Flanker task. However, although these tools are digital and free of charge, prior knowledge is required to select the tests to be applied, and they do not present information regarding the norms for each population.

On the other hand, tests that assess more general skills are used in educational contexts, which sometimes include the assessment of executive functions or some of their components. This is the case for tests such as the Woodcock-Muñoz battery (Muñoz-Sandoval et al. 2005) and the WISC-V test (Rosas et al. 2022), which include specific components related to the assessment of executive functions. Finally, and especially in school contexts, some scales assess executive functions indirectly, in different contexts, and through the appreciation of actors close to the children, such as teachers or relatives. Among the most widely used are the Behavior Rating Inventory of Executive Function (BRIEF, Gioia et al. 2000), the Behavior Assessment System for Children, now in its third edition (BASC, Reynolds and Kamphaus 2015), and the Conners test (Conners 2008).

However, the assessment systems mentioned above present certain limitations because, on the one hand, the tests used in the research area assess executive functions in a general way without detailing aspects related to their components. On the other hand, the assessment of executive functions in school contexts only considers executive functions as a minor aspect of more general skills, such as cognitive ability. Moreover, the scales that focus on the appreciation of third parties tend to mark a tendency towards the less cognitive aspects of executive functions, generating a biased view of their development. On the other hand, there are doubts about the validity of these instruments, which are discussed in the next section.

1.4. Discussion of the Importance of Direct EF Assessment over Indirect Ones

Executive functions are important for children's behavior and learning, but what method is best for assessing these abilities? Much research shows low correlations between the results of direct and indirect assessments of executive functions. In a review of 20 studies reporting correlations between the two types of measures, Toplak et al. (2013) found that only 24% of all reported correlations were statistically significant and that the median correlations were only r = 0.19 (equating to only 3.6% common variance). It should be noted that this result cannot be attributed to the lack of reliability or validity of both methods since both indirect and direct scales showed quite good psychometric properties. So, how can two types of assessment that are supposed to measure the same thing have such low correlations? A recent study sought to answer this question. Even though both measurement forms can show good predictive abilities for academic performance (Gerst et al. 2015), the evidence seems to indicate that direct cognitive tests are more efficient and robust than indirect assessments for measuring executive functioning. The study by Soto et al. (2020), conducted with 136 children, clearly showed how executive function assessments made by teachers adequately predicted students' academic assessments (also made by teachers) but failed to predict academic performance. Tests of executive functions instead predict academic performance very well and predict academic ratings even better than indirect assessments of executive functions.

This study is of particular relevance since, to date, it is the only published research with two independent and two dependent variables and, in both cases, with direct and indirect methods. Moreover, this makes it possible to elucidate more precisely what both techniques measure; the academic assessments seem to better measure better school adjustment according to teachers, while the direct ones are a better measure of school adjustment and academic achievement. Thus, it would appear that direct measures are more accurate and would be a better indicator of executive functioning than indirect measures.

1.5. Brief Description of the Yellow-Red Battery

The Yellow-Red battery consists of six tests focused on the general assessment of executive functions and the specific assessment of their different components. The assessment system is based on technological support (Tablet) and is within the paradigm of invisible assessment through play (Rosas et al. 2015). The test was designed to be applied to children aged 6 to 11 years and has a total application time ranging from 15 to 30 min.

1.5.1. Cat-Dog

The first test, called Cat-Dog, is an adaptation of Diamond's Hearts and Flowers test (Diamond 2013). This test theoretically measures the three components of executive functions in its three phases. In the first congruent phase, participants must touch the same side of the screen where a stimulus (cat) appears. The second phase is incongruent: participants must press the opposite side of the screen to where the stimulus appears (dog). In each of the first two phases, 12 cats or dogs appear. In the third phase, congruent and incongruent stimuli (cats and dogs) appear randomly 33 times. In all phases, the stimuli are displayed for 1 s with an interval of 500 milliseconds. Points are only awarded for the results obtained in the third phase. One point is assigned for each correct response, and 0 points are assigned for omissions and incorrect or anticipatory responses, i.e., those executed by the participant before 200 milliseconds elapse. As seen below, this test theoretically evaluates the flexibility component of executive functions (Figure 2).

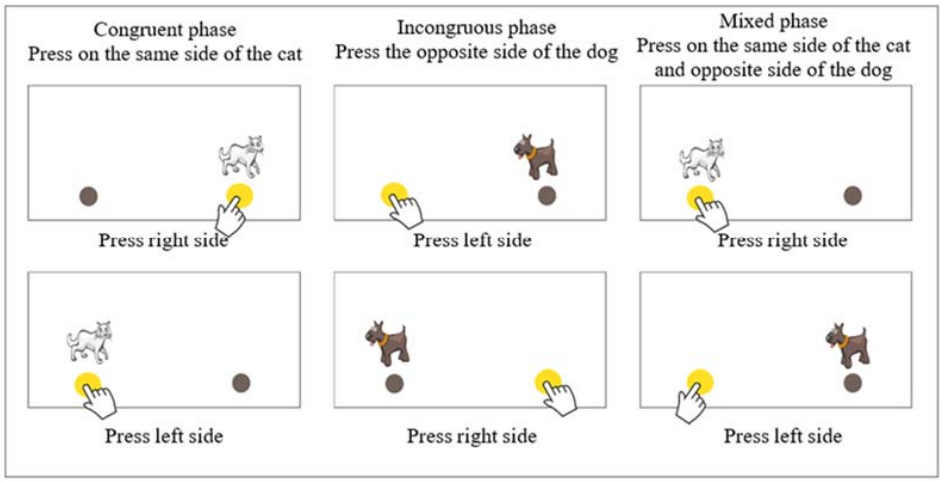

Figure 2. Description of the phases of the Cat-Dog test.

1.5.2. Arrows

This test evaluates cognitive inhibition and attention; a "model" arrow and three arrows that function as response alternatives appear on the screen. The arrows point to the right, to the left, up, or down. In the first three cases, children must press the arrow pointing in the same direction as the model. However, the participants should not press anything when the arrow points downward. This test has 36 items, 8 of which correspond to inhibition tasks. The first 15 items are displayed for 2 s, with 500-millisecond intervals, while the following 21 items are presented for 1 s, with 500-millisecond intervals. One point is awarded for each correct response, and 0 points for incorrect or anticipatory response (response with a reaction time less than 200 milliseconds) (Figure 3).

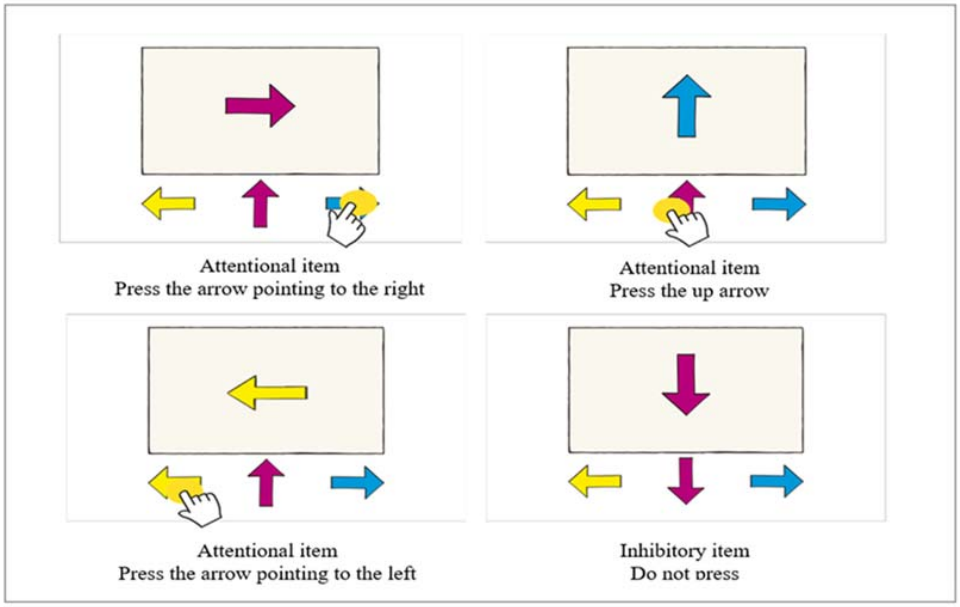

Figure 3. Description of the items of Arrows test.

1.5.3. Flies

The Flies test assesses behavioral inhibition using a delay of gratification. A screen is presented with flies flying in different directions, and the participant is asked to smash as many as possible. The flies make a buzzing sound as they fly, and when smashed, they make a sound that the children find very amusing and rewarding. When a green light is turned on, the participant can continue smashing the flies; however, when the light turns red, the participant should not continue smashing the flies. The traffic light changes color,

and the participant must follow the rule. When it is green, you can smash flies; when it is red, you cannot.

The test lasts 2 min and is divided into eight different time-lapses where the red or green light appears. Each time-lapse lasts between 3 and 10 s. One point is awarded for each fly smashed. The delayed gratification indicator is the sum of the flies smashed in the green minus those smashed in the red-light time lapses (Figure 4).

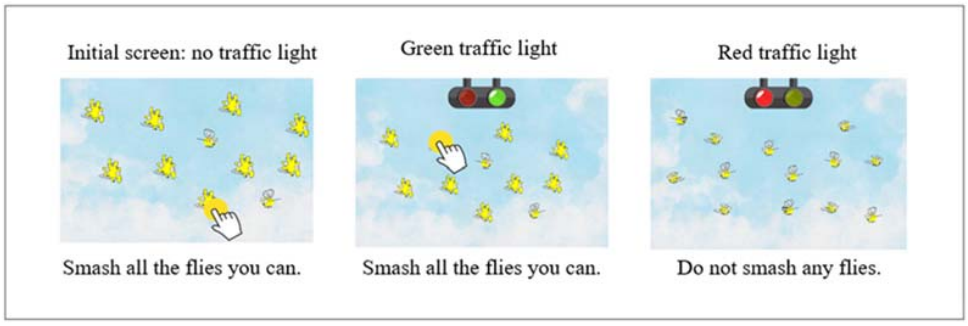

Figure 4. Description of the Flies test.

1.5.4. Binding

This test evaluates the development of visuospatial working memory in the form of associated pairs. A series of images related to numbers or geometric figures are presented. Then, some of the stimuli are presented again in isolation, and the participants must establish the associations according to how they were initially presented. The test has 27 items; as the test progresses, more images and numbers are added. In the case of the youngest children (6 to 8 years old), the first five items use geometric figures instead of numbers. From age nine onwards, only items with numbers are presented, and 0 points are assigned for each incorrect answer. An answer in which all pairs are appropriately associated is considered correct; if there is at least one mistake, the item is considered incorrect (Figure 5).

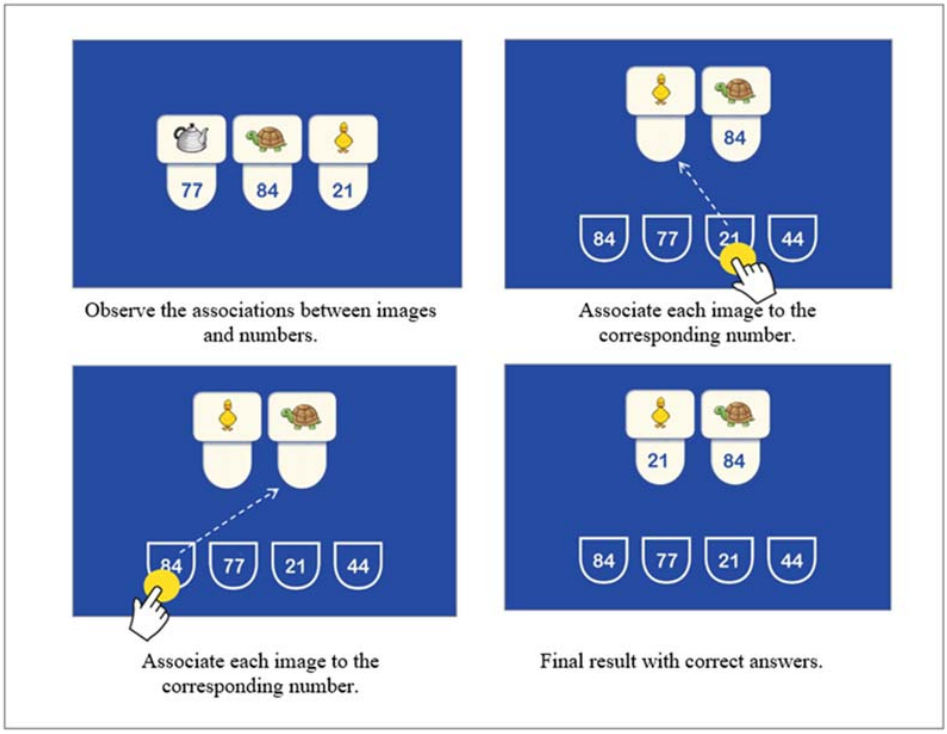

Figure 5. Example of Binding test item.

1.5.5. The Farm

This test evaluates auditory and visual working memory. The evaluation of auditory working memory is performed by presenting a sequence of animal sounds, after which the participant must select the corresponding animals on a board starting from the last sound heard (Figure 6).

Figure 6. Example of Farm auditory test item.

A keyboard is displayed on which some keys light up to assess visual working memory. The participant must press the keys in the reverse order in which they are illuminated.

The auditory sequences range from 2 to 8 sounds, and the visual sequences from 2 to 10 visual stimuli. There are 18 auditory items and 18 visual items. One point is assigned for each correct answer. The test is failed when two consecutive errors are made at the same level (the level is determined by the number of sounds to be remembered by the participant) (Figure 7).

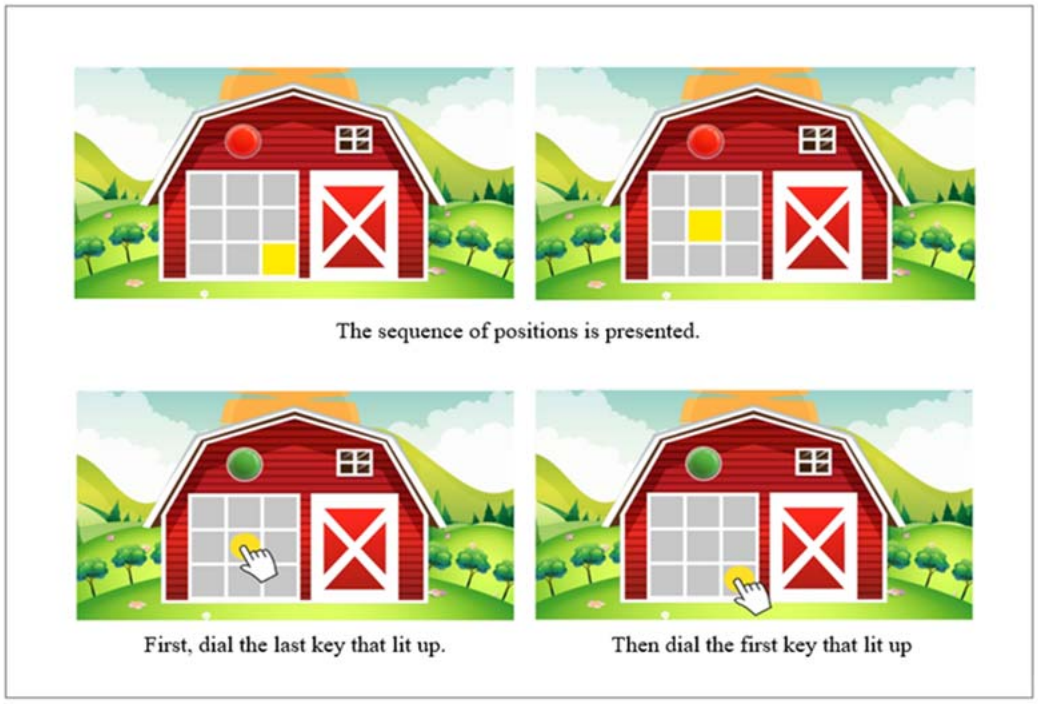

Figure 7. Example of Farm visual test item.

1.5.6. Triads

The Triads test is oriented to evaluate cognitive flexibility. A series of four geometric figures are presented, three of which have a common characteristic (color, shape, or size). Participants must choose three that have something in common, but the classification criteria are not made explicit. These implicit criteria are color, shape, and size. The criteria change without giving any warning. In total, the test has 21 items, 5 of which correspond to the implicit criterion of color, 5 to the implicit criterion of shape, and 5 to the implicit criterion of size. Six random criterion items follow this. Participants have three chances to get it right; if they fail, they skip to the next category, and those omitted items are considered. Each failed attempt is considered an attentional error, but if it fails all three attempts, it is considered a perseverative error. One point is obtained for each correct answer in the first attempt; in the second attempt, 0.6 points, and in the third attempt, 0.3 points. One point is deducted for each attentional error and 2 points for each perseverative error. There is no time limit for the permanence of the items. This test is suspended after three incorrect answers (Figure 8).

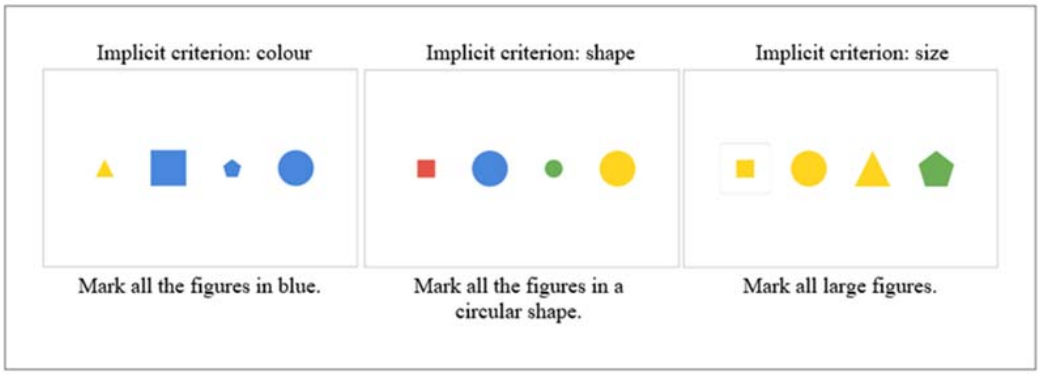

Figure 8. Description of the *Triads* test levels.

In accordance with the evidence reviewed and the presentation of the instrument developed and standardized to assess executive functions in school-aged children, this study has the following objectives: firstly, at a general level, to demonstrate the importance of having direct standardized measures of executive functions that can be used with children of a wide age range; secondly, to obtain the psychometric properties of the Yellow-Red Test. Finally, we sought to clarify the factor structure of executive functions in Chilean children between 6 and 12 years of age.

2. Methodology

The Chilean standardization of Yellow-Red had a meticulous design to have enough information to validate the test with the gold standards for each of the components of executive functions: inhibition, working memory, and flexibility. These gold standards provide valuable information regarding the evidence of validity with other variables and the validity of an instrument developed under the stealth assessment paradigm (Rosas et al. 2015).

2.1. Instruments

A table summarizing the correspondence between the Yellow-Red subtests and the respective gold standards applied in the study is presented in Table 1.

Table 1. Correspondence of Yellow-Red subtests with their gold standards.

General Component	Specific Component	Yellow-Red Subtest	Gold Standard
Inhibition	Cognitive inhibition and attention	Arrows	Flankers
	Behavioral inhibition	Flies	Flankers
Working Memory	Visuospatial working memory	Binding	Digit Span WISC V
	Auditory and visual working memory	Farm	Digit Span WISC V
			Block Design WISC V
Cognitive Flexibility		Triads	Modified Card Sort Test
			Similarities WISC V

2.1.1. Hearts and Flowers

This test assesses executive functions in general and specifically, according to the assessment block. For the present research, only the third phase was used, which assesses cognitive flexibility (A. Diamond, personal communication, April 2018). The Inquisit Web 6 platform (Millisecond Software LLC 2021), which allows offline tablet assessment, was used. Specifically, the Chilean Spanish language version was programmed as described by the original authors of the instrument (Borchert 2021; Diamond et al. 2007). Participants see a set of items in which a heart or a flower appears on the right or left side of a fixation cross. If the person sees a heart, he or she must press a "button" on the tablet on the same side as the heart, which is called a congruent item. On the other hand, if the person sees a flower to the right or left of the fixation cross, he/she must press the button on the opposite side from where the flower appears, which is called an incongruent item. The test consists of three blocks: the first is congruent, in which 20 items (hearts only) are presented, with ten random appearances on each side of the fixation cross. According to Diamond et al. (2007), the congruent block evaluates working memory. The incongruent block also has 20 items, this time only with flowers, with ten flowers on the left and ten flowers on the right of the fixation cross. According to the authors, this block evaluates working memory in addition to inhibitory control. Congruent and incongruent blocks have a maximum response time of 5000 ms. The last block is mixed, and in it, participants must respond to 20 congruent and incongruent items that appear randomly, which assesses working memory, inhibitory control, and cognitive flexibility. The maximum response time in this block is 6000 ms. Each block has three practice items, for which automatic feedback is given to the participant, and failure is a criterion for suspension from the test. For the present study, and due to the differences in presentation times that facilitate a response in the Hearts and Flowers test, compared to Cat-Dog, those correct responses selected 1000 ms after stimulus presentation were scored with 0.5 points (equivalent to half of the correct response). In addition, responses selected before 200 ms scored 0 points.

2.1.2. Flankers

This test evaluates attention and cognitive inhibition. The Inquisit Web 6 platform (Millisecond Software LLC 2021) was used. The test corresponds to the Chilean Spanish version and follows the procedure described by the authors of the instrument (Borchert 2021; Rueda et al. 2004). The test consists of presenting an image of five fish lined up. The participant must pay attention to the fish in the center, and if the fish in the center looks to the right, the participant must press the button on the right. However, if the fish in the center faces left, the participant must press the key on the left. The other fishes in the row can look in the same direction as the fish in the center (congruent item) or in the opposite direction to the fish in the center (incongruent item) (see Figure 9). The platform presents two practice blocks, 12 items in which only four fishes are presented (6 looking left and six looking right), and 12 events with five fishes (three compatible events looking left, three compatible events looking right, three incompatible events looking left, and three incompatible events looking right). The maximum response time per item is 3000 ms. For the present study, and due to differences in presentation times that facilitate

responding in the Flankers test compared to Yellow-Red tests that assess inhibition, those correct responses selected after 1500 ms following stimulus presentation were scored with 0.5 points (equivalent to half of the correct response). In addition, responses selected before 200 ms were scored with 0 points.

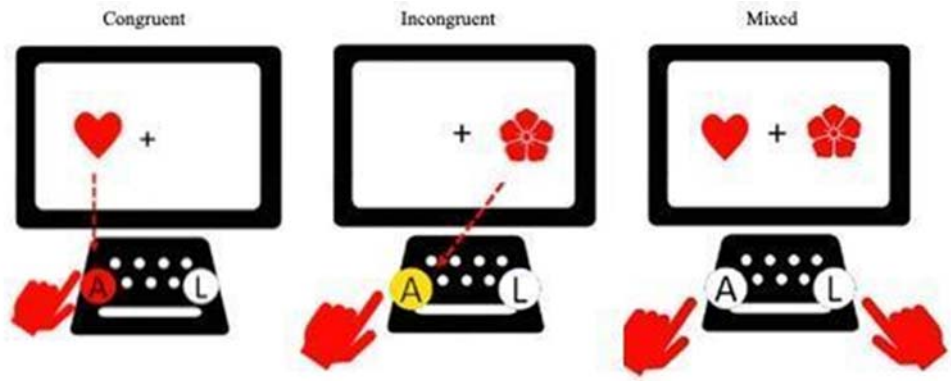

Figure 9. Description of the Hearts and Flowers test phases.

2.1.3. Modified Card Sort Test

This test assesses cognitive flexibility in the face of changes in the rules of the task. It is a child version of the Wisconsin Card Sorting Test, in which fewer items are considered than in the original version, and only unambiguous cards are presented for sorting. What is essential is the ability to search for a new sorting category when the rule changes implicitly. The Inquisit Web 6 platform (Millisecond Software LLC 2021) was used. The test corresponds to the Chilean Spanish version and follows the procedure described by the authors of the instrument (Borchert 2021; Nelson 1976). The test consists of the subject having to classify one letter per item according to the similarity in a category with one of the four letters displayed below. Two blocks of 24 items are presented, and each of the 24 cards presented in each block has a maximum of one characteristic in common with the four response cards; thus, there are no ambiguous items. The cards in each block are presented randomly and without repetition. For each block, the same order of sorting criteria is followed: color, shape, quantity, repeating the pattern consecutively (see Figure 10). The rule changes automatically after six consecutive correct answers in each category. The score corresponds to the number of correct answers (Figure 11).

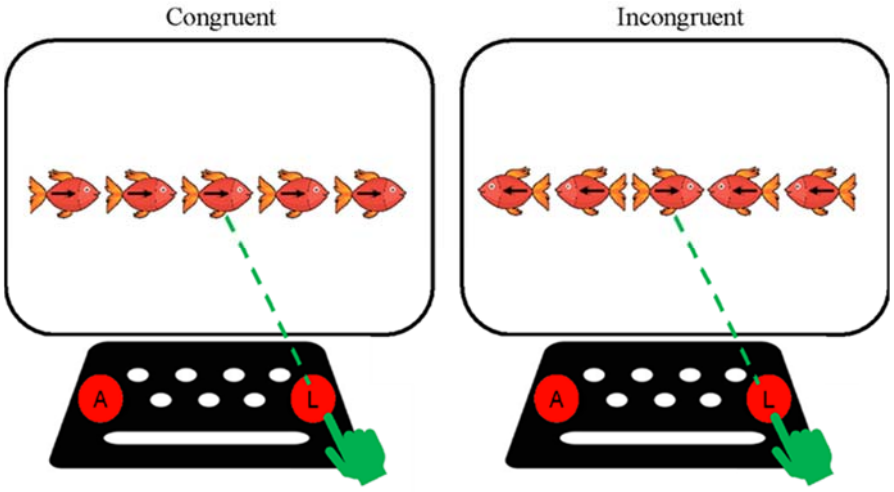

Figure 10. Description of test items *3.

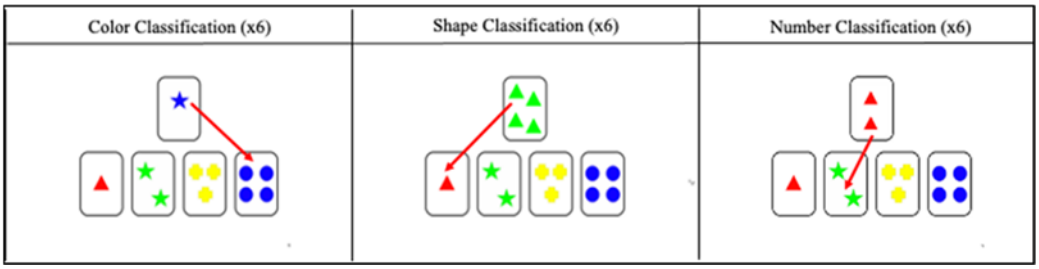

Figure 11. Description of classification categories of the Modified Card Sort Test.

2.1.4. Digit Span

This test corresponds to a subtest of the Wechsler Intelligence Scale for Children, 5a edition (WISC-V), in its standardized version for the Chilean population (Rosas et al. 2022). This subtest corresponds to one of the two tests that make up the working memory index (Rosas and Pizarro 2018). It comprises three tasks: (1) Digits in Direct Order: a sequence of numbers is read to the participant, which he or she must repeat in the same order. (2) Digits in Reverse Order: the second sequence of numbers is read to the participant, which the person must repeat in reverse order. (3) Sequenced Digits: in the third sequence of digits, the participant must repeat them in ascending order. Each task includes two practice attempts to ensure the understanding of the task. Each of the three tasks has nine items, each containing two attempts. As the items progress, they have more items to remember; for example, item 1 has numbers, while item 9 has ten numbers in the Digits in Direct Order task. The test has a suspension criterion, which is applied when the child makes a mistake in two attempts at the same item. The total score corresponds to the sum per task of each correct attempt (one point) answered by the subject.

2.1.5. Verbal Fluency Test

Researchers decided to develop a task to assess verbal cognitive flexibility. The test consisted of naming as many items that met specific characteristics as possible in 60 s. The instruction given to the children was the following: "Please name as many things that you like as fast as you can". The answers were recorded to be scored later, considering as a score the sum of words expressed within the time limit. This test is not part of the Yellow-Red battery; it was developed to have a second test of verbal flexibility to contribute to the factorial structure of the model.

2.1.6. School Adaptation Index

This index corresponds to a composite score between the TRF and the average grades obtained by the student the year prior to the evaluation. The index is expressed in percentiles, in which each measure weighs 50%.

Teacher Report Form 6–18, Spanish version (TRF): To obtain information on school adjustment, four questions were used that are not part of the questionnaire itself but of a section of contextual questions on general aspects of the adjustment observed by the educator, such as the degree to which the educator perceives that the student makes an effort, if the student seems happy, if he/she behaves appropriately, etc. The questions were answered in Likert format, with zero points indicating "Much less than the average of their peers," while a score of seven indicated "Much less than the average of their peers" (Achenbach and Rescorla 2001). Thus, the maximum possible raw score was 28 points. This score was transformed into a percentile so that a score of 28 points was equivalent to 100% TRF-Adaptation.

Academic Performance: The second measure of school adjustment corresponds to the grade point average obtained by the student in 2020. This average ranged from 0 to 7 and was reported by the students' schools. The average obtained was transformed to a percentile so that grade seven corresponded to 100% academic performance.

2.2. Sample

The Yellow-Red standardization process was approved by the Scientific Ethical Committee of Social Sciences, Arts, and Humanities of the Pontificia Universidad Católica de Chile. Sampling focused on subjects having sufficient variability in the following characteristics: SES, gender, age, and school adjustment, for which the individual school adjustment index described in the previous section was used. Table 2 shows the sample distribution according to SES, gender, and level of education of each student.

Table 2. Sample distribution by SES, gender, and grade.

	Low SES				**High SES**					
	Female		**Male**		**Female**		**Male**		**Total**	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Kinder	9	3.67	9	3.67	12	4.90	14	5.71	44	17.96
1st Grade	10	4.08	11	4.49	10	4.08	10	4.08	41	16.73
2nd Grade	10	4.08	10	4.08	10	4.08	10	4.08	40	16.33
3rd Grade	3	1.22	7	2.86	12	4.90	12	4.90	34	13.88
4th Grade	9	3.67	12	4.90	12	4.90	9	3.67	42	17.14
5th Grade	10	4.08	7	2.86	10	4.08	10	4.08	37	15.10
6th Grade	2	0.82	1	0.41	3	1.22	1	0.41	7	2.86
Total		21.6		23.2		28.1		26.9	24	100.0

Note. SES = Socioeconomic Status.

The sample consisted of 245 participants (122 girls); 110 belonged to the low SES and 135 to the high SES. The socioeconomic categorization proposed by the Quality Agency of the schools attended by the students was considered to define their SES. This categorization considers the parents' educational level, the family's average monthly income, and the vulnerability index. This index was calculated based on the percentage of students in the school who are in extreme poverty or at risk of school failure. The first three indicators were obtained through a survey answered by the families of the children who took the SIMCE test (a national standardized test to assess Math and Language). The last one was obtained from the Junta Nacional de Auxilio Escolar y Becas (JUNAEB) data.

The students came from schools in Santiago, Chile, in grades from kindergarten to 6th grade.

All students took all the tests. Only two children did not participate in the Flies subtest. These missing cases are due to one child with suspected color blindness (necessary for the test) and one child not responding.

The principal of each school agreed to participate in the study and gave the authorization to contact students' families. All participants were authorized by their parents or legal guardians through a letter of informed consent; they also went through the informed consent process.

2.3. Procedure

Students were assessed at home or at their educational establishments during school hours in a room provided by each school to carry out the procedure. The evaluations were carried out at the end of the current school year (second semester 2021). The assessment was conducted in one 60-min session for students in grades two through six and two 30-min sessions for those in kindergarten and first grade. A trained evaluator administered all tests individually in oral or digital format (Tablet format).

Traditional format tests were administered first, i.e., oral question-answer tests. These were the subtests of the WISC-V. In this case, the evaluators recorded the answers on the Tablet, and after the evaluation, they scored the tests. Subsequently, the tests were applied in digital format; first were the gold standard tests, Hearts and Flowers, followed by the Flanker test and the Card Sort Task, which were presented randomly, according to the definition in the programming of the application. Then, the Yellow-Red battery tests were

applied in the following order: Cat-dog, Triads, Arrows, Binding, Farm (auditory and visual), and Flies. Finally, the verbal fluency test was applied.

Regarding missing data, the analyses were carried out considering the two students mentioned above as missing cases and using the listwise method for the multivariate analyses.

The data analysis was performed with SPSS 27 for all analyses except the confirmatory factor analysis performed with Mplus 8. The analysis plan was structured as follows: to determine internal consistency, a reliability test was performed. To assess the validity of Yellow-Red, the progression of scores according to age was evaluated, using the course attended by the students as a proxy for this variable. The difference in scores for each test between kindergarten, third, and fifth grade was reported. To obtain evidence of convergent and discriminant validity of the Yellow-Red test, an analysis of correlations with reference variables (gold standards) and with a variable that theoretically did not correlate with the instrument was carried out. To check the structure of the test, three confirmatory factor analyses were carried out: for the full sample, for younger students, and for older students.

3. Results

3.1. Evidence of Reliability

As can be seen in Table 3, the evidence of test reliability is excellent, both in internal consistency indicators and in bipartition indicators (for the Flies test, an internal consistency indicator could not be calculated). In five of six tests, the coefficients exceeded 0.8, which is considered very good. In the Cat-Dog test, the value was above 0.9, which is considered excellent.

Table 3. Reliability analysis, internal consistency, or bipartition indicator.

Yellow Red	Cronbach's Alpha	Pearson's r
Flies		0.82 **
Arrows	0.88	
CatDog	0.91	
Binding	0.86	
Farm	0.82	
Triads	0.86	

Note: Interpretation of Cronbach's alpha values: $\alpha \geq 0.9$, excellent; $0.9 > \alpha \geq 0.8$, good; $0.8 > \alpha \geq 0.7$, acceptable; $0.7 > \alpha \geq 0.6$, questionable; $0.6 > \alpha \geq 0.5$, poor; $0.5 > \alpha$, unacceptable. ** = $p < 0.01$.

3.2. Evidence of Validity

3.2.1. Progression of Executive Functions with Age

The progression of the results according to age showed clear evidence of the tests' validity. As shown in Figure 12, all the tests, except for Farm, showed an evident progression concerning age. Although, in all of them, a flattening of the curves was observed towards older ages, which may indicate that the test decreased its discriminative capacity as the age of the children increases. Regarding Farm, it reached its highest value in third grade and practically did not increase in the following grades (Figure 12).

A one-way ANOVA was performed to test the progression of scores according to age. The test was conducted considering kindergarten, second, and fifth grade, since the sample size decreased in sixth grade. As can be seen in Table 4, all tests showed a statistically significant increase in scores according to age. Additionally, the effect sizes were all above 0.14, which is considered large. The post hoc Tukey analysis indicated that the difference of 1.47 points (95% CI [−3.28, 0.34] $p = 0.136$) on average between kindergarten and second grade on the Farm test was the only comparison that was not statistically significant.

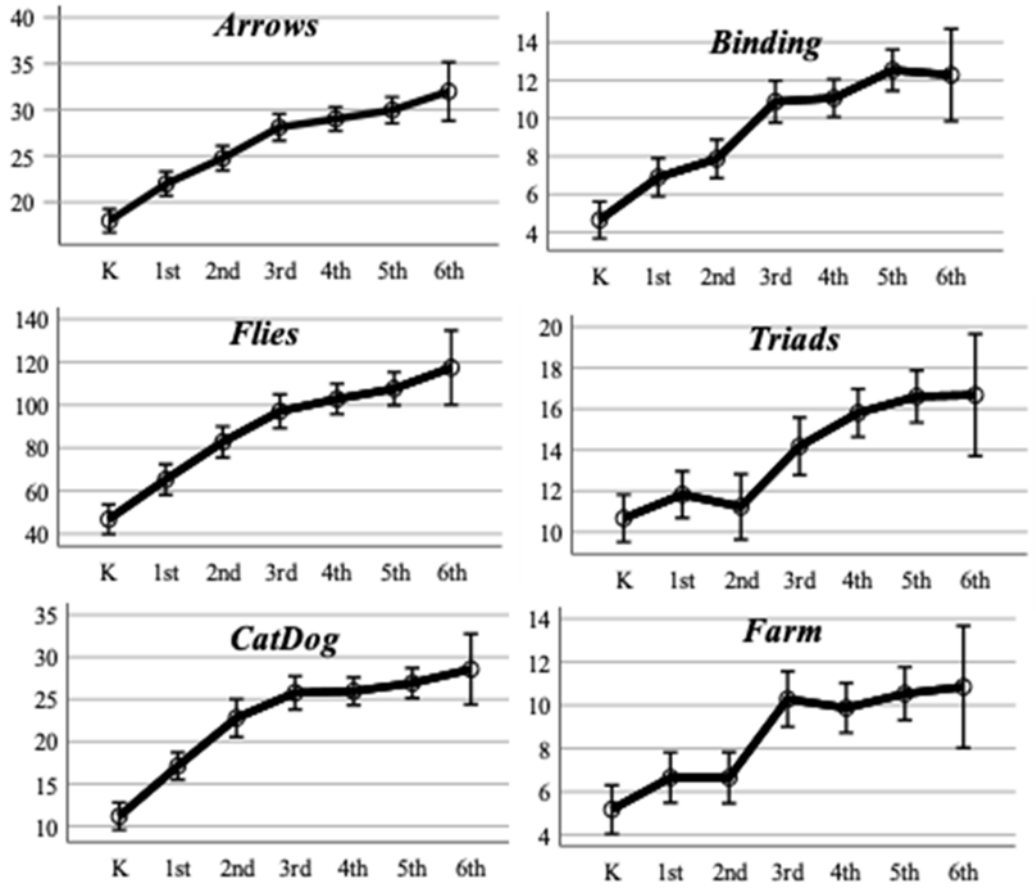

Figure 12. Age progression of estimated marginal means by subtest.

Table 4. ANOVA results.

		N	M	SD	SE	*F*(2,118)	η^2
Arrows	Kinder	44	18.000	4.393	0.662	87.773 ***	0.598
	2nd grade	40	24.775	4.554	0.720		
	5th grade	37	30.108	3.204	0.527		
Binding	Kinder	44	4.659	2.828	0.426	55.644 ***	0.485
	2nd grade	40	7.875	3.750	0.593		
	5th grade	37	12.568	3.516	0.578		
Flies	Kinder	44	46.818	24.819	3.742	68.352 ***	0.541
	2nd grade	40	82.825	24.142	3.817		
	5th grade	35	107.629	20.150	3.406		
Triads	Kinder	44	10.611	2.639	0.398	28.096 ***	0.323
	2nd grade	40	12.810	4.501	0.712		
	5th grade	37	16.600	3.498	0.575		
CatDog	Kinder	44	11.386	5.809	0.876	88.948 ***	0.601
	2nd grade	40	22.063	5.330	0.843		
	5th grade	37	26.500	4.531	0.745		
Farm	Kinder	44	5.182	3.208	0.484	24.789 ***	0.296
	2nd grade	40	6.650	3.393	0.537		
	5th grade	37	10.541	3.884	0.639		

*** $p < 0.001$.

3.2.2. Evidence of Convergent and Discriminant Validity

Table 5 presents the correlations of the YR tests with their corresponding gold standards and with a test that theoretically should not correlate significantly with executive

functions, e.g., the mental health variable as perceived by teachers on the TRF scale. This variable corresponds to the response to the question to what degree is (the student) happy and content? As can be seen in Table 5, all the correlations observed were significant at 1%, with values ranging from 0.42 (flexibility) to 0.73 (behavioral inhibition). Likewise, all the correlations with the discriminant test were close to zero and insignificant.

Table 5. Correlations between Yellow-Red subtests and their corresponding gold standard and TRF (mental health test, not related to measured skills).

Yellow-Red Subtest	The Gold Standard (Reference)	Correlation	Correlation with TRF
Flies	Flankers	0.59 **	0.02
Arrows	Flankers	0.72 **	0.03
CatDog	Hearts and flowers	0.63 **	0.00
Binding	Digit Span WISC V	0.57 **	0.08
Farm	Digit Span WISC V	0.63 **	0.06
Triads	Modified Card Sort Test	0.42 **	0.12
	Oral Fluency	0.37 **	0.17*

Note WISC V Digits: Mean score obtained in forward, backward, and sequencing digits. Correlation using Pearson's r: ** = $p < 0.01$.

3.2.3. Evidence of Factorial Validity

The factorial structure of executive functions during childhood and early adolescent development goes from being a unitary construct to being progressively differentiated into two factors. Later in adolescence and adulthood, it shows a structure in which three differentiated factors emerge but maintain a certain level of correlation between them (Lehto et al. 2003; Shing et al. 2010; Willoughby et al. 2012). In the present study, we sought to test whether the Yellow-Red subtests also reflected the increasing diversification of the components of executive functions with age. For this, a series of CFAs were carried out, which allowed for interpreting the relationship between observed variables (in this case, the results in each of the Yellow-Red tests) with latent variables or the components of the executive functions, and the relationship existing between the latent variables. The three latent variables defined are: inhibition (I), for which the observed variables corresponded to the results of the Arrows (ARR) and Flies (FLI) subtests. The latent variable working memory (WM) was composed of the observed variables Binding (BIN) and Farm (FAR), while the latent variable cognitive flexibility (CF) was composed of Triads (TRI) and Cat-Dog (CAT). The models tested ranged from most differentiated to least differentiated. Thus, the first model evaluated a structure of three latent factors. The second model was composed of two factors, in which two components were combined into one latent variable and a third factor was left with only one latent variable. This model had three versions, in order to test all possible combinations between components. Finally, the simplest model was the one with a single latent variable, which grouped all the components of executive functions (see Table 6).

Table 6. Proposed models for confirmatory factor analysis.

Model	Factor 1	Factor 2	Factor 3
Model 1	I	WM	CF
Model 2A	(I + WM)	CF	
Model 2B	I	(WM + CF)	
Model 2C	(I + CF)	WM	
Model 3	(I + WM + CF)		

Note. I = inhibition, WM = working memory, CF = cognitive flexibility.

To test the increasing differentiation in the components of executive functions, the analysis was carried out in three groups; one considered the entire sample, i.e., students from kindergarten to sixth grade. The second group considered the youngest children in the study, kindergarten and first grade students. Finally, the analysis was conducted with the older children, students from second to sixth grade. The CFA was performed using a robust maximum likelihood estimator (MLM), which allows the analysis to be performed with variables that present distributions with a certain level of abnormality. For each group analyzed, a table is presented with the five models and their respective adjustment statistics. The model with the best fit to the data was highlighted, which was then represented graphically, indicating the parameters for each of the variables.

Each of these models was tested under three conditions: the first condition considered the whole sample, the second condition considered only the youngest children in the sample (kindergarten and first grade), and the third condition included only the oldest students (second to sixth grade).

Confirmatory factor analysis: Kindergarten to sixth grade

The results shown in Table 7 indicate that, for the whole sample, the model that best fit the data corresponded to model 3 (see Figure 13), which considered a single factor that included the three components of executive functions. Models 1, 2.A, and 2.B presented latent factors with correlations greater than 1, which prevented an accurate estimation of their degree of fit. In this case, it was recommended to collapse the factors that presented problems. Models 2C and 3 presented statistically significant χ^2 values and *p*-values, indicating a low level of fit. However, this could be due to the small number of participants, as the χ^2 statistic is very sensitive to the N of the sample (Byrne 2011). When analyzing the rest of the goodness-of-fit indicators for models 2C (CFI = 0.974; RMSEA = 0.095; SRMR = 0.038) and 3 (CFI = 0.974; RMSEA = 0.089; SRMR = 0.38), we saw that both models presented adequate values for CFI (greater than 0.95) and SRMR (lower than 0.08), but presented difficulties concerning RMSEA (values higher than 0.06 were observed). However, model 3 presented a value closer to what was expected; thus, it was considered the most adequate model.

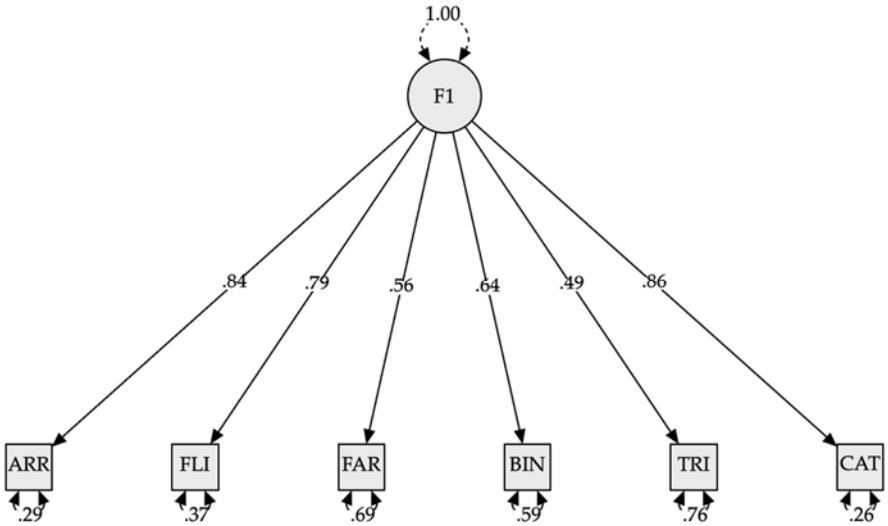

Figure 13. Model plot confirmatory factor analysis (kindergarten to sixth grade). *Note.* F1 = Model 3 (I + WM + CF). ARR = subtest Arrow. FLI = subtest Flies. FAR = subtest Farm. BIN = subtest Binding. TRI = subtest Triads. CAT = subtest Cat-Dog. All modeled correlations and path coefficients are statistically significant ($p < 0.001$).

Table 7. Goodness of fit indices for alternative CFA models (kindergarten to sixth grade).

Model	*df*	χ^2	*p*	CFI	RMSEA	RMSEA 95% CI	SRMR
(1) I-WM-CF [a]	6	19.545	0.003	0.980	0.096	0.051–0.146	0.033
(2.A) (I + WM)-CF [a]	8	19.883	0.010	0.982	0.078	0.035–0.122	0.033
(2.B) I-(WM + CF) [a]	8	25.013	0.002	0.975	0.094	0.054–0.136	0.039
(2.C) (I + CF)-WM	8	25.624	0.001	0.974	0.095	0.055–0.138	0.038
(3) (I + WM + CF)	**9**	**26.308**	**0.002**	**0.974**	**0.089**	**0.051–0.129**	**0.038**

Note. [a] = The latent variable covariance matrix (PSI) is not positive definite. The model with the best fit is highlighted in bold.

Confirmatory factor analysis: Kindergarten and first grade

As we can see in Table 8, the confirmatory factor analysis results for kindergarten and grade 1 children indicated that the best-fitting model was model 3 (see Figure 14), which indicates the existence of a single latent factor grouping the observed variables. The CFI value of 0.92 indicated an adequate fit. However, RMSEA (greater than 0.10) and SRMR (greater than 0.065) indicated a low model fit. However, this could be due to the small sample size. Models 1 and 2A, 2B, and 2C were discarded because the latent variable covariance matrix (PSI) was not positive definite.

Table 8. Goodness of fit indices for alternative CFA models (kindergarten and first grade).

Model	*df*	χ^2	*p*	CFI	RMSEA	RMSEA 95% CI	SRMR
(1) I-WM-CF [a]	6	7.772	0.255	0.985	0.059	0.000–0.161	0.041
(2.A) (I + WM)-CF [a]	8	13.765	0.088	0.950	0.092	0.000–0.172	0.055
(2.B) I-(WM + CF) [a]	8	18.551	0.018	0.909	0.125	0.049–0.200	0.065
(2.C) (I + CF)-WM [a]	8	14.601	0.067	0.943	0.099	0.000–0.1770	0.056
(3) (I + WM + CF)	**9**	**18.731**	**0.028**	**0.916**	**0.113**	**0.036–0.185**	**0.065**

Note. [a] = The latent variable covariance matrix (PSI) is not positive definite. The model with the best fit is highlighted in bold.

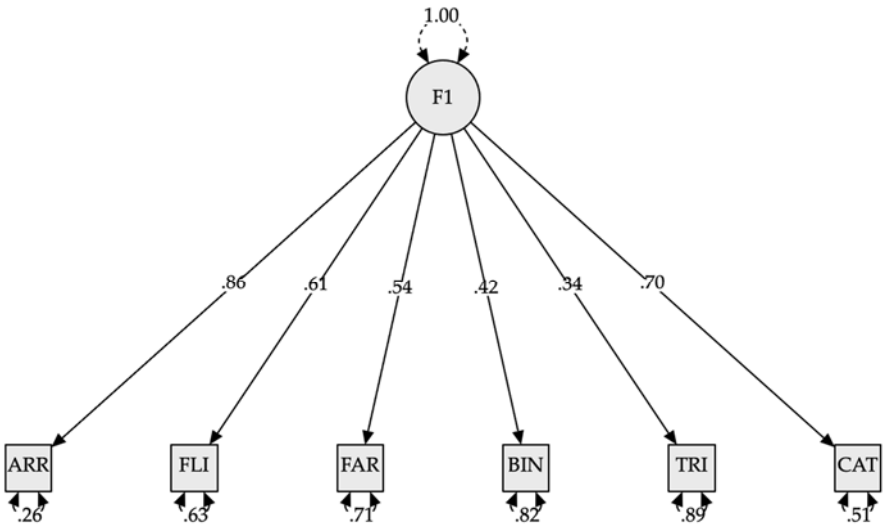

Figure 14. Model plot confirmatory factor analysis (kindergarten and first grade). *Note.* F1 = Model 3 (I + WM + CF). ARR = subtest Arrows. FLI = subtest Flies. FAR = subtest Farm. BIN = subtest Binding. TRI = subtest Triads. CAT = subtest Cat-Dog. All modeled correlations and path coefficients are statistically significant ($p < 0.001$).

Confirmatory factor analysis: Second to sixth grade

The best fitting model for students in grades 2–6 was model 2C (CFI = 0.926; RMSEA = 0.095; SRMR = 0.52). Models 1, 2A, and 2B were unacceptable, as the latent variables had correlations greater than 1. Model 3 was discarded as it had a lower fit than model 2C (see Table 9). Model 2-C is represented in Figure 15.

Table 9. Goodness of fit indices for alternative CFA models (second to sixth grade).

Model	*df*	χ^2	*p*	CFI	RMSEA	RMSEA 95% CI	SRMR
(1) I-WM-CF [a]	6	16.714	0.010	0.931	0.106	0.047–0.168	0.048
(2.A) (I + WM)-CF [a]	8	21.337	0.006	0.913	0.103	0.051–0.157	0.053
(2.B) I-(WM + CF) [a]	8	22.321	0.004	0.907	0.106	0.055–0.160	0.056
(2.C) (I + CF)-WM	**8**	**19.474**	**0.012**	**0.926**	**0.095**	**0.042–0.150**	**0.052**
(3) (I + WM + CF)	9	23.917	0.004	0.903	0.102	0.054–0.153	0.056

Note. [a] = The latent variable covariance matrix (PSI) is not positive definite. The model with the best fit is highlighted in bold.

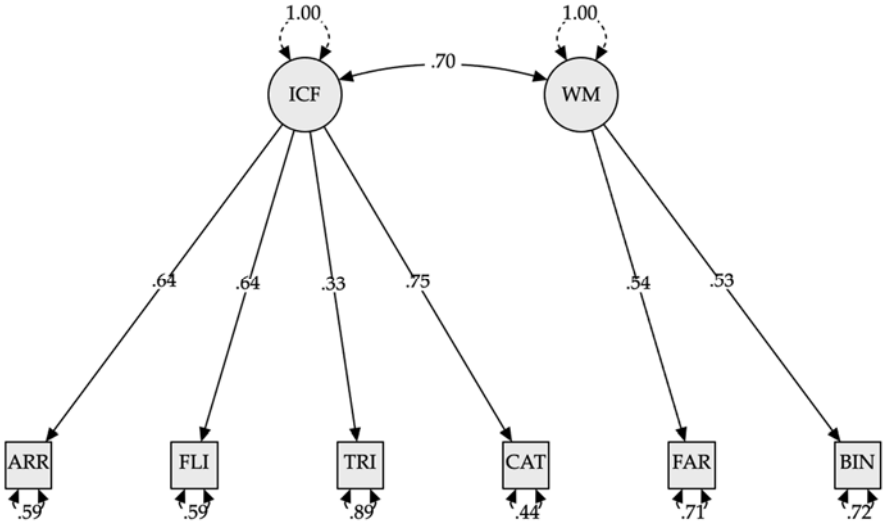

Figure 15. Model plot confirmatory factor analysis (second to sixth grade). *Note*. ICF = Inhibition plus cognitive flexibility, WM = working memory, ARR = Arrows, FI = Flies, TRI = Triads, CAT = Cat-Dog, FAR = Farm, BIN = Binding. All modeled correlations and path coefficients are statistically significant ($p < 0.001$).

4. Discussion

This article described the Yellow-Red test's standardization process for assessing executive functions in children aged 6 to 11. The development of this test responded to the need for instruments that allow valid and reliable measurement of executive functions due to the high impact they have on academic performance, as well as on work adaptation and emotional stability in adulthood (Diamond 2016).

The need for instruments designed specifically for child assessment, which also have solid validity and reliability indicators, was highlighted. The Yellow-Red test adequately responded to this demand, being, to our knowledge, the only instrument to measure executive functions playfully, based on a Tablet format, which independently assesses the three basic components of executive functions postulated by Miyake et al. (2000) and has psychometric evidence that supports its reliability and validity.

Yellow-red is a test designed specifically for assessing executive functions in children, and its development considered the four aspects raised by Germine et al. (2019). The Yellow-Red test has several advantages over other assessment instruments. The test is attractive to

users because, on the one hand, the use of a digital medium that can be manipulated directly by children takes into account the high technological proficiency demonstrated by students belonging to generations of digital natives (Mcmanis and Gunnewig 2012; Sweeney and Geer 2008). On the other hand, it incorporates playful dynamics, which has been shown to positively impact children's motivation and engagement (Perrotta et al. 2013). On the other hand, using playful tests with digital support increases the chances of accuracy in the assessment results, especially in the case of children with learning difficulties (Rosas et al. 2015). Finally, as there is no evidence that the digital format interferes negatively with the cognitive assessment process of children (Kingston 2009; Wang et al. 2008), the incorporation of technology and gamification, such as the Yellow-Red battery, is considered relevant.

Although there are other instruments designed to assess the various components of executive functions, both independently and in general, many of these do not have information regarding their psychometric properties, which significantly reduces their reliability. Examples of this are Espy's Shape School test (Espy 1997), which assesses flexibility and inhibition in preschoolers; the Dimensional Change Card Sorting Test (DCCS) (Zelazo et al. 1996), which assesses flexibility; and the Hearts and Flowers test, which theoretically assesses the three components, initially called Dots (Davidson et al. 2006; Diamond et al. 2007).

On the other hand, they are standardized instruments with evidence of reliability and validity. However, they are oriented only to one of the components of executive functions, or they frame their evaluation as a part of a more general cognitive function. This is the case for the WISC-V Digit Span subtest, which assesses working memory in isolation based on the application of three subtests (digit span forward, backward, and sequencing), of which strictly only the last two assess working memory (the first assesses short term memory); for the Woodcock-Muñoz number reversal and auditory working memory subtests, which are oriented to the assessment of working memory; and for the concept formation subtest, related to cognitive flexibility (Schrank et al. 2005). Thus, although we have standardized instruments, their design was not focused on evaluating executive functions and even less on the specific assessment of their components. Still, they are oriented to general cognitive skills or school performance.

However, other instruments have several characteristics attributed to the Yellow-Red battery. An example of this is the battery developed by Zelazo et al. (2013), which is also game-based and presented in a Tablet format, and it proposes the evaluation of the three basic components of executive functions; a clear differentiation of the components cannot be made because the total assessment is based only on the use of two tests (DCCS and Flankers), which could generate a contamination of the tasks and therefore difficulty in isolating the performance in each of its components. On the other hand, Yellow-Red has at least one test to evaluate each component, which strengthens the possibilities of differentiation.

One of the main pieces of evidence of the validity of the Yellow-Red test is the excellent correlations obtained with instruments considered gold standards for each component. The convergent and discriminant validity analysis showed conclusive results regarding the quality of each subtest to assess, respectively, the factors of inhibition, working memory, and flexibility.

On the other hand, test results generally show a progression with age. That is, performance improves in direct relation to age. This is in line with the progressive development of executive functions, which has been widely described in the literature (Friedman and Miyake 2017; Miyake et al. 2000; Gathercole 1998). As shown, the onset of executive function development begins during the first months of age and continues into adulthood. The results presented here align with the expected progress in executive function development. One particularly striking result is the performance in the 'Farm' test (Figure 12). Here, a clear difference was seen between the results obtained by participants in grades K-2 and grades 3–6. This result may be related to the progressive development of working memory,

which increases by one unit of information every two years. Thus, given that there are participants of different ages in each grade, the K–2 group corresponds to participants with an average age of 7.2 years (StdDev = 0.9), and the 2–6 grade group corresponds to participants with an average age of 10.4 years (StdDev = 0.9). A possible explanation for this observed result is that, in the second group (3–6 grades), the development of some skills necessary for a good performance in WM tests has started (Gathercole 1998). In particular, rehearsal and practice-by-repetition strategies, whose development starts at seven years, should be more present in the second group (3rd–6th grade) (Chooi and Logie 2020; Morra 2015).

The confirmatory factor analysis of the six Yellow-Red subtests, plus the one carried out with the gold standard tests, recognizes and reaffirms that Chilean children's executive function structures progressively differentiate with age. Starting with a common factor at preschool age, as demonstrated by Wiebe et al. (2008) and Willoughby et al. (2012), until around six years of age and differentiating into two factors in middle childhood, from seven to 12 years of age, as found by Shing et al. (2010). One factor linked to inhibition and another related to cognitive flexibility, sharing working memory tasks, can be distinguished. However, as Wiebe (2014) points out, the evidence for two factors is robust, but their composition is not.

Future research should clarify whether, by taking a wider age range, the three factors can be more clearly differentiated. For this, the test should be applied to students up to at least 12 years of age.

5. Conclusions

After a rigorous development, adaptation, and standardization process, it was possible to develop a battery for an exhaustive assessment of executive functions, considering global indicators and specific measures for each primary component. All the subtests that compose it have sufficient evidence of validity and reliability. The Yellow-Red executive function assessment battery responds to the needs of the academic, clinical, and educational communities. By having an appropriate assessment instrument, researchers and practitioners can accurately assess children's executive functions at an early age, which allows for the detection of possible difficulties and the strengthening of skills that could present difficulties.

Future research will explore the longitudinal use of this tool in various contexts and demonstrate the practical impact of a playful, technology-based instrument that allows both the general and isolated assessment of the different components of executive functions. This could imply the consideration of these results in the design of clinical interventions and educational programs, which, considering the broad impact of executive functions in diverse areas of human development, could contribute to the possibilities of correcting difficulties and strengthening those aspects necessary to improve people's quality of life.

6. Limitations

The present study showed strong evidence for the validity and reliability of Yellow-Red. However, the sample size comprised rather small groups, mainly when classified by age and SES. These sample sizes may affect the interpretability and extrapolation of results to more diverse populations. In this sense, complementing the present study with studies conducted in other populations, countries, and cultures would provide information on the instrument's usability in other cultural contexts.

Author Contributions: Conceptualization, R.R. and V.E.; methodology, C.S.-C. and C.M.; validation, C.M., V.E. and R.R.; formal analysis, C.S.-C. and C.M.; investigation, C.S.-C., R.R. and V.E.; resources, C.S.-C.; data curation, C.S.-C.; writing—original draft preparation, R.R., V.E., C.S.-C. and C.M.; writing—review and editing, R.R. and V.E.; visualization, C.M.; supervision, R.R.; project administration, C.S.-C.; funding acquisition, R.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ANID PIE CIE160007.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics committee of Social Sciences, Arts and Humanities of Pontificia Universidad Católica de Chile (protocol code 201230006 and date 21 April 2021 of approval) for studies involving humans.

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 FigShare at https://doi.org/10.6084/m9.figshare.21701687.v1.

Acknowledgments: We want to thank the international collaboration in developing this test. Special thanks to Adele Diamond and Klaus Oberauer, authors of the Cat-Dog and Binding tests. To Paul Collard and Diane Fischer-Naylor, for the English version and for using the test to validate the intervention programs of Creativity, Culture and Education, England. To Per Norman Anderson, for the Norwegian version. Annemarie Fritz, Lars Orbach and Jörg Hampe, for the German version. Szilvia Németh, for the Hungarian version. Raluca Ciulei, for the Rumanian version. Bingqing Tang, for the Chinese version. And finally, to Bernardita Ovalle, for her great help with the references.

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

References

Achenbach, Thomas M., and Leslie A. Rescorla. 2001. *Manual for the ASEBA School-Age Forms and Profiles*. Burlington: ASEBA.

Baddeley, Alan. D., and Graham J. Hitch. 1994. Developments in the Concept of Working Memory. *Neuropsychology* 8: 485–93. [CrossRef]

Borchert, Katja. 2021. *User Manual: Inquisit Hearts and Flowers Task (Chilean Spanish version)*. Seattle: Milliseconds.

Byrne, Barbara M. 2011. *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming*. Oxfordshire: Routledge. [CrossRef]

Cameron Ponitz, Claire E., Megan M. McClelland, Abigail M. Jewkes, Carol McDonald Connor, Carrie L. Farris, and Frederick J. Morrison. 2008. Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Research Quarterly* 23: 141–58. [CrossRef]

Camerota, Marie, Michael T. Willoughby, Brooke E. Magnus, and Clancy B. Blair. 2020. Leveraging item accuracy and reaction time to improve measurement of child executive function ability. *Psychological Assessment* 32: 1118–32. [CrossRef] [PubMed]

Canavan, Anthony, Ivan Janota, and Paul H. Schurrt. 1985. Luria's frontal lobe syndrome: Psychological and anatomical considerations. *Neurosurgery, and Psychiatry* 48: 1049–53. [CrossRef]

Carlson, Stephanie M. 2005. Developmentally Sensitive Measures of Executive Function in Preschool Children. *Developmental Neuropsychology* 28: 595–616. [CrossRef]

Chaudron, Stephane, Rosanna Di Gioia, and Monica Gemo. 2018. *Young Children (0-8) and Digital Technology: A Qualitative Study across Europe*. Luxembourg: Publications Office. [CrossRef]

Chooi, Weng-Tink, and Robert Logie. 2020. Changes in error patterns during n-back training indicate reliance on subvocal rehearsal. *Memory & Cognition* 48: 1484–503.

Conners, C. Keith. 2008. *Conners 3*. North Tonawanda: MHS.

Cowan, Nelson. 2017. The many faces of working memory and short-term storage. *Psychonomic Bulletin and Review* 24: 1158–70. [CrossRef]

Davidson, Matthew C., Dima Amso, Loren Cruess Anderson, and Adele Diamond. 2006. Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia* 44: 2037–78. [CrossRef]

Day, Jamin, Kate Freiberg, Alan Hayes, and Ross Homel. 2019. Towards Scalable, Integrative Assessment of Children's Self-Regulatory Capabilities: New Applications of Digital Technology. *Clinical Child and Family Psychology Review* 22: 90–103. [CrossRef]

Diamond, Adele. 2013. Executive functions. In *Annual Review of Psychology*. Palo Alto: Annual Reviews Inc., vol. 64, pp. 135–68. [CrossRef]

Diamond, Adele. 2016. Why improving and assessing executive functions early in life is critical. In *Executive Function in Preschool-Age Children: Integrating Measurement, Neurodevelopment, and Translational Research*. Washington, DC: American Psychological Association, pp. 11–43. [CrossRef]

Diamond, Adele, W. Steven Barnett, Jessica Thomas, and Sarah Munro. 2007. Preschool Program Improves Cognitive Control. *Science* 318: 1387–88. [CrossRef] [PubMed]

Espy, Kimberly A. 1997. The shape school: Assessing executive function in preschool children. *Developmental Neuropsychology* 13: 495–99. [CrossRef]

Friedman, Naomi. P., and Akira Miyake. 2017. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. In *Cortex*. Milano: Masson SpA, vol. 86, pp. 186–204. [CrossRef]

Garon, Nancy, Susan E. Bryson, and Isabel M. Smith. 2008. Executive Function in Preschoolers: A Review Using an Integrative Framework. *Psychological Bulletin* 134: 31–60. [CrossRef]

Gathercole, Susan E. 1998. The development of memory. *The Journal of Child Psychology and Psychiatry and Allied Disciplines* 39: 3–27. [CrossRef]

Germine, Laura, Katharina Reinecke, and Naomi S. Chaytor. 2019. Digital neuropsychology: Challenges and opportunities at the intersection of science and software. *The Clinical Neuropsychologist* 33: 271–86. [CrossRef] [PubMed]

Gerst, Elyssa H., Paul T. Cirino, Jack M. Fletcher, and Hanako Yoshida. 2015. Cognitive and behavioral rating measures of executive function as predictors of academic outcomes in children. *Child Neuropsychology* 23: 381–407. [CrossRef]

Gioia, Gerard A., Peter K. Isquith, Steven C. Guy, and Lauren Kenworthy. 2000. *Behavior rating inventory of executive function: BRIEF*. Odessa: Psychological Assessment Resources.

Golden, Charles J. 1978. *Stroop Color and Word Test: A Manual for Clinical and Experimental Uses*. Wood Dale: Stoelting.

Homack, Susan, and Cynthia A. Riccio. 2004. A meta-analysis of the sensitivity and specificity of the Stroop Color and Word Test with children. *Archives of Clinical Neuropsychology* 19: 725–43. [CrossRef]

Hughes, Claire, and Rosie Ensor. 2011. Individual differences in growth in executive function across the transition to school predict externalizing and internalizing behaviors and self-perceived academic success at 6 years of age. *Journal of Experimental Child Psychology* 108: 663–76. [CrossRef]

Kane, Michael J., and Randall W. Engle. 2003. Working-Memory Capacity and the Control of Attention: The Contributions of Gol Neglect, Response Competition, and Task Set to Stroop Interference. *Journal of Experimental Psychology: General* 132: 47–70. [CrossRef]

Kingston, Neal M. 2009. Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education* 22: 22–37. [CrossRef]

Lee, Kerry, Rebecca Bull, and Ringo M. H. Ho. 2013. Developmental changes in executive functioning. *Child Development* 84: 1933–53. [CrossRef] [PubMed]

Lehto, Juhani E., Petri Juujärvi, Libbe Kooistra, and Lea Pulkkinen. 2003. Dimensions of executive functioning: Evidence from children. *British Journal of Developmental Psychology* 21: 59–80. [CrossRef]

Mcmanis, Lilla Dale, and Susan B. Gunnewig. 2012. Finding the Education in Educational Technology with Early Learners. *Young Children* 67: 14–24.

Millisecond Software LLC. 2021. *Inquisit Web 6*. Seattle: Millisecond Software LLC.

Miyake, Akira, Naomi P. Friedman, Michael J. Emerson, Alexander H. Witzki, Amy Howerter, and Tor D. Wager. 2000. The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology* 41: 49–100. [CrossRef]

Moran, Lisa, and Keith Owen Yeates. 2018. Stroop Color and Word Test, Children's Version. In *Encyclopedia of Clinical Neuropsychology*. Edited by Jeffrey S. Kreutzer, John DeLuca and Bruce Caplan. Cham: Springer International Publishing, pp. 3323–25.

Morra, Sergio. 2015. How do subvocal rehearsal and general attentional resources contribute to verbal short-term memory span? *Frontiers in Psychology* 6: 145. [CrossRef] [PubMed]

Muñoz-Sandoval, Ana F., Richard W. Woodcock, Kevin S. McGrew, and Nancy Mather. 2005. *Batería III Woodcock-Muñoz: Pruebas de habilidades cognitivas*. Rolling Meadows: Riverside Publishing Company.

Nelson, Hazel E. 1976. A Modified Card Sorting Test Sensitive to Frontal Lobe Defects. *Cortex* 12: 313–24. [CrossRef]

Parsey, Carolyn. M., and Maureen Schmitter-Edgecombe. 2013. Applications of technology in neuropsychological assessment. *Clinical Neuropsychologist* 27: 1328–61. [CrossRef]

Perrotta, Carlo, Gill Featherstone, Helen Aston, and Emily Houghton. 2013. *Game-Based Learning: Latest Evidence and Future Directions*. Slough: NFER (National Foundation for Educational Research).

Plaisted, James R., Greta N. Wilkening, John L. Gustavson, and Charles J. Golden. 1983. The Luria-Nebraska Neuropsychological Battery—Children's Revision: Theory and Current Research Findings. *Journal of Clinical Child Psychology* 12: 13–22. [CrossRef]

Portellano, José Antonio, M. Rosario Martínez, and Lucía Zumárraga. 2011. *Evaluación Neropsicológica de las funciones ejecutivas en niños*. Madrid: TEA Ediciones.

Reynolds, Cecil R., and Randy W. Kamphaus. 2015. *BASC3: Behavior Assessment System for Children*. San Antonio: PsychCorp.

Rosas, Ricardo, and Marcelo Pizarro. 2018. *WISC-V. Manual de Administración y Corrección*. Macul: CEDETi-UC.

Rosas, Ricardo, Francisco Ceric, Andrés Aparicio, Paulina Arango, Rodrigo Arroyo, Catalina Benavente, Pablo Escobar, Polín Olguín, Marcelo Pizarro, María P. Ramírez, and et al. 2015. ¿Pruebas tradicionales o evaluación invisible a través del juego? Nuevas fronteras de la evaluación cognitiva. *Psykhe* 24: 1–11. [CrossRef]

Rosas, Ricardo, Marcelo Pizarro, Olivia Grez, Valentina Navarro, Dolly Tapia, Susana Arancibia, María Teresa Muñoz-Quezada, Boris Lucero, Claudia P. Pérez-Salas, Karen Oliva, and et al. 2022. Estandarización Chilena de la Escala Wechsler de Inteligencia para Niños - Quinta Edición. *Psykhe (Santiago)* 31: 1–23. [CrossRef]

Rozenblatt, Shahal. 2018. Stroop Color Word Test (Adult). In *Encyclopedia of Clinical Neuropsychology*. Edited by Jeffrey S. Kreutzer, John DeLuca and Bruce Caplan. Cham: Springer International Publishing, pp. 3325–27. [CrossRef]

Rueda, M. Rosario, Jin Fan, Bruce D. McCandliss, Jessica D. Halparin, Dana B. Gruber, Lisha Pappert Lercari, and Michael I. Posner. 2004. Development of attentional networks in childhood. *Neuropsychologia* 42: 1029–40. [CrossRef] [PubMed]

Schrank, Fredrick, Kevin McGrew, Mary L. Ruef, Criselda G. Alvarado, Ana F. Muñoz-Sandoval, and Richard W. Woodcock. 2005. *Overview and Technical Supplement (Batería III Woodcock-Muñoz Assessment Service Bulletin No. 1)*. Rolling Meadows: Riverside Publishing.

Shing, Yee L., Ulman Lindenberger, Adele Diamond, Shu-Chen Li, and Matthew C. Davidson. 2010. Memory Maintenance and Inhibitory Control Differentiate From Early Childhood to Adolescence. *Developmental Neuropsychology* 35: 679–97. [CrossRef] [PubMed]

Soto, Elia F., Michael J. Kofler, Leah J. Singh, Erica L. Wells, Lauren N. Irwin, Nicole B. Groves, and Caroline E. Miller. 2020. Executive functioning rating scales: Ecologically valid or construct invalid? *Neuropsychology* 34: 605–19. [CrossRef] [PubMed]

Strommen, Ellen A. 1973. Verbal Self-Regulation in a Children's Game: Impulsive Errors on "Simon Says". *Child Development* 44: 849–53. [CrossRef]

Stroop, John Ridley. 1935. Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology* XVIII: 643. [CrossRef]

Subsecretaría de Telecomunicaciones de Chile. 2017. *IX Encuesta de Acceso y Usos de Internet*. Lo Prado: Subsecretaría de Telecomunicaciones de Chile, pp. 1–66.

Sweeney, Trudy, and Ruth Geer. 2008. Student capabilities and attitudes towards ICT in the early years (Student capabilities and attitudes towards ICT in the early years). *Australian Educational Computing* 25: 18.

Toplak, Maggie E., Richard F. West, and Keith E. Stanovich. 2013. Practitioner Review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry and Allied Disciplines* 54: 131–43. [CrossRef]

United States Census Bureau. 2021. Computer and Internet Use in the United States: 2018. Available online: https://www.census.gov/newsroom/press-releases/2021/computer-internet-use.html (accessed on 11 October 2022).

Wang, Shudong, Hong Jiao, Michael J. Young, Thomas Brooks, and John Olson. 2008. Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement* 68: 5–24. [CrossRef]

Wiebe, Sandra A. 2014. Modeling the emergent executive: Implications for the structure and development of executive function. *Monographs Society Res Child* 79: 104–15. [CrossRef]

Wiebe, Sandra A., Kimberly Andrews Espy, and David Charak. 2008. Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. *Developmental Psychology* 44: 575–87. [CrossRef]

Willoughby, Michael T., Clancy B. Blair, R. J. Wirth, and Mark Greenberg. 2012. The measurement of executive function at age 5: Psychometric properties and relationship to academic achievement. *Psychological Assessment* 24: 226–39. [CrossRef] [PubMed]

Zelazo, Philip D., Clancy B. Blair, Michael T. Willoughby, Meredith Larson, Erin Higgins, and Amy Sussman. 2016. Executive Function: Implications for Education. Available online: https://ies.ed.gov/ncer/pubs/20172000/pdf/20172000.pdf (accessed on 7 December 2022).

Zelazo, Philip D., Douglas Frye, and Tanja Rapus. 1996. An age-related dissociation between knowing rules and using them. *Cognitive Development* 11: 37–63. [CrossRef]

Zelazo, Philip D., Jacob E. Anderson, Jennifer Richler, Kathleen Wallner-Allen, Jennifer L. Beaumont, and Sandra Weintraub. 2013. II. NIH Toolbox Cognition Battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development* 78: 16–33. [CrossRef] [PubMed]

Journal of
Intelligence

MDPI

Perspective

Clinical Reasoning: A Missing Piece for Improving Evidence-Based Assessment in Psychology

Gabrielle Wilcox *, Meadow Schroeder and Michelle A. Drefs

School and Applied Child Psychology, Werklund School of Education, University of Calgary, Calgary, AB T2N 1N4, Canada
* Correspondence: gwilcox@ucalgary.ca

Abstract: Clinical reasoning is a foundational component of conducting evidence-based psychological assessments. In spite of its importance, limited attention has been paid to the teaching or measurement of clinical reasoning skills relative to psychological assessment, as well as how clinical reasoning develops or how its efficacy can be measured. Improving clinical reasoning throughout the assessment process, from initial case conceptualization to hypotheses testing, to recommendation writing, has the potential to address commonly noted concerns regarding diagnostic accuracy, as well as the accessibility and utility of psychological reports and recommendations, and will, ultimately, lead to improved outcomes for clients. Consequently, we provide a definition of clinical reasoning in relation to psychological assessment, followed by a critique of graduate training assessment and the current challenges of measuring clinical reasoning in psychology. Lastly, this paper provides suggestions for how to incorporate clinical reasoning throughout the assessment process as a way to answer client questions more effectively and provide meaningful recommendations to improve outcomes.

Keywords: clinical reasoning; critical thinking; evidence-based assessment

Citation: Wilcox, Gabrielle, Meadow Schroeder, and Michelle A. Drefs. 2023. Clinical Reasoning: A Missing Piece for Improving Evidence-Based Assessment in Psychology. *Journal of Intelligence* 11: 26. https://doi.org/10.3390/jintelligence11020026

Received: 11 October 2022
Revised: 13 January 2023
Accepted: 19 January 2023
Published: 26 January 2023

1. Introduction

Evidence-based assessment (EBA) is a relatively new concept in psychology that emphasizes the theory and research in selecting and using high-quality assessment methods and processes (Youngstrom and Van Meter 2016). Although there are no agreed-upon standards for its application in psychology, there have been some attempts at providing guidelines for EBA, based on the American Psychological Association's (American Psychological Association 2006) three recommendations for evidence-based psychological practice, including: (a) using the best available research, (b) applying clinical expertise, and (c) attending to patient characteristics, culture, and preferences (Bornstein 2017). Others have noted that EBA requires effective critical thinking and reasoning, which informs all aspects of assessment, from determining the questions and choosing assessment measures to interpreting the results by analyzing information and data within the context of a client (Dombrowski et al. 2021; Victor-Chmil 2013; Ward 2019). Thus, clinical reasoning supports clinicians who must engage in clinical reasoning during assessment and make diagnostic decisions when presenting client problems in EBA.

The purpose of this paper is to describe the current state of clinical reasoning research in the context of psychological assessment and to propose potential directions for promoting clinical reasoning in assessment practice. This paper will first define the role of clinical reasoning in evidence-based assessment and the research related to this area, outlining some of the contemporary challenges in the training and research related to clinical reasoning in assessment. The second section will summarize the current, albeit limited, literature on how psychologists develop clinical reasoning skills, along with recommendations for extending the research findings on deliberate practice (DP). Finally, this paper will suggest how practitioners might be able to improve their clinical reasoning in assessment contexts, based on the findings of medicine and psychotherapy.

2. The Role of Clinical Reasoning in Evidence-Based ()Assessment

Victor-Chmil (2013) posited that "*critical thinking* is the cognitive processes used for analyzing knowledge" (p. 34) and also that "*clinical reasoning* is the cognitive and metacognitive processes for analyzing knowledge relative to a clinical situation or specific patient" (Victor-Chmil 2013, p. 34). Often used interchangeably with other terms such as critical reasoning, clinical reasoning allows psychologists to make sense of a large amount of data as they develop working hypotheses, identify information that supports or refutes those hypotheses, and compare data to diagnostic criteria. Both critical reasoning and clinical reasoning involve intentionally thinking about a problem, testing hypotheses, and generating solutions to the problem (American Psychological Association n.d.; Gruppen 2017). Critical thinking requires not only attending to the outcome of the process but also attending to the process of thinking, which is often omitted in research on assessment (Gambrill 2019). Because clinical reasoning and critical reasoning are notoriously poorly or inconsistently defined within the literature, and because there is considerable overlap between these two terms, they are considered similar enough that we have used clinical reasoning in this paper, due to its more common use within the broader research literature.

In order to move toward EBA and utilize clinical expertise in this process, it is important to understand the current challenges of implementing EBA (Ward 2019). One challenge is in understanding how clinicians gain and apply the foundational skill of clinical reasoning in psychological assessment (Dombrowski et al. 2021). Reasoning is an under-discussed topic in EBA (Wright et al. 2022) that is used when testing hypotheses related to clients' functioning within their context, synthesizing and integrating data from multiple sources, and providing diagnoses and meaningful treatment recommendations to improve functioning (Mash and Hunsley 2005; Wright et al. 2022; Youngstrom et al. 2015; Youngstrom and Van Meter 2016). When performed well, clinical reasoning aids psychologists in asking important questions to ensure that consideration is given to how psychologists' beliefs about clients or their problems influence the assessment process.

Unfortunately, faulty clinical reasoning can lead to misdiagnoses and may harm clients through delayed, insufficient, or inappropriate treatment, which ultimately leads to a lack of faith in psychological services (Gambrill 2012; Wright 2021). Currently, there are no available statistics on how faulty clinical reasoning affects the general population because of the difficulty in directly connecting error rates in psychology to negative outcomes (Gambrill 2012). This contrasts with the medical field, where there are considerably more publications on this topic owing to the availability within the medical field of more objective measures of error rates, such as mortality and the length of hospital stays (e.g., Ahmed et al. 2015). Specific to psychology, the link between poor critical reasoning and negative client outcomes is largely indirect and has primarily been examined in relation to the common types and sources of errors in both the testing and report-writing processes, largely ignoring the role of critical reasoning in these problems. Because of the important role that psychological assessment can play in improving client functioning, understanding how psychologists think and reason critically throughout the process of assessment and case conceptualization is vital for improving the quality of assessments (Siegert 1999). Additionally, while there have been significant advances in evidence-based treatments, the lack of corresponding attention to EBA is surprising, as treatment selection should be informed by assessment (Mash and Hunsley 2005).

The pursuit of clinical reasoning in assessment is an important goal. The conclusions and diagnostic decisions derived from psychoeducational assessment can have a significant effect on the daily lives of clients. For instance, an understanding of the ecological factors that either support or restrict the success of a student with academic difficulties is critical in determining whether or not the student meets the diagnostic criteria for a learning disability, and identifying the appropriate remediation, learning support at home and school, and accommodations that are specific to that pupil's educational needs.

3. Examining the Current State of Research Training, Research, and Practice

As Gordon et al. (2022) aptly remarked, "Clinical reasoning is a topic that often feels familiar (or even obvious) . . . [however,] this sense of familiarity may be masking important differences in how it is understood, operationalized, and assessed" (p. 109). Indeed, how psychologists engage in clinical reasoning during assessment has largely been neglected in the literature (Mash and Hunsley 2005). In discussing the current state of clinical reasoning in psychology, we have drawn upon research into the technical aspects of test administration (Oak et al. 2019), the use of base rates (Burns 1990), diagnostic accuracy in assessment (Aspel et al. 1998; de Mesquita 1992; Watkins 2009), and improving report writing (Nelson 2021; Pelco et al. 2009; Postal et al. 2018), but this body of literature had not addressed how to develop and improve critical reasoning in psychological assessment. Much of the argument of this paper is based on the research of clinical reasoning skills in social work (Gambrill 2012, 2019), medicine (Young et al. 2020), and psychological therapy (Miller et al. 2020) because clinical reasoning and how to develop and improve clinical reasoning in psychological assessment has largely been ignored. Below, we review what is known about clinical reasoning from the literature, highlighting issues with how it is taught, researched, and, currently, practiced.

3.1. A Focus on Testing Rather than Assessment

One of the challenges in understanding the role of clinical reasoning in assessment has been the commonplace conflation of the terms, "testing" and "assessment". In training assessment skills, an emphasis on standardized assessment and reducing administrative error in training programs is warranted, as standardized administration requires considerable training, and critical thinking is predicated on quality data. However, paying attention to testing, including choosing appropriate measures with strong psychometric properties and interpreting test scores appropriately, is imperative but it is insufficient to ensure strong clinical reasoning. Testing generally refers to choosing and administrating measures and assessment alignments. Assessment, however, refers to the entire process, from choosing what questions to ask during the initial interview to interpreting all of the data gathered, including but not limited to test scores (Canivez 2019; Suhr 2015; Wright 2021); the initial steps inform the subsequent hypotheses and guide the assessment process, but they occur prior to test selection, administration, and interpretation (Ward 2019). One problem with most evaluations of assessment skills in training is that there is an emphasis on evaluating the psychometric aspects of assessment and standardized test administration, at the expense of clinical reasoning development (Mash and Hunsley 2005; Wright 2021). There is a danger in focusing on the generation of test scores at the expense of clinical reasoning. Psychologists can use psychometrically strong measures and administer them appropriately but will come to poor conclusions if they do not have the clinical reasoning skills to determine what the problem is that is being presented, in order to ask and answer the right questions or to integrate and interpret the resulting data effectively (Mash and Hunsley 2005).

During the psychological assessment process, test scores are an important source of information. Learning the standardized test measures is a complex and time-consuming task that represents an important foundational skill for reducing error and increasing reliability. Error is inherent in testing for various reasons such as client and examiner factors, as well as problematic testing conditions, including incomplete data, time pressures, and complex environments; therefore, it is important to reduce administrative error as much as possible. Unfortunately, despite the focus on a standardized assessment, errors are common. For example, despite the fact that these are learned skills that are a core part of training programs, assessment errors are commonplace, with practitioners often making more errors than students (Oak et al. 2019). This level of difficulty in accurately implementing skills that are essential for assessment contributes to poor clinical reasoning by providing poor-quality data.

3.2. Test-by-Test Reporting

The concern that emphasizing test scores over assessment can lead to weak clinical reasoning is demonstrated by the dominant test-by-test approach used in report writing, which some argue reflects the quality of clinical reasoning (Pelco et al. 2009). It is important that reports are transparent when explaining how the psychologist arrived at their diagnostic conclusions, along with how the assessment process informed the diagnostic decision and recommendations, but test-by-test reports do not make psychologists' reasoning transparent (Pelco et al. 2009; Wilcox and Schroeder 2015). Weak clinical reasoning can contribute to unclear reports that do not support the clients. In this regard, errors in both the assessment and report-writing processes provide indirect evidence of the association between poor clinical reasoning and negative client outcomes.

Along these lines, Wright (2021) has cogently described the current state of clinical reasoning in assessment: "Psychological assessment has long been a mysterious, intuited process, taught to psychologists in training, test by test, with components of conceptualization, integration, and report writing somewhat tacked onto the end of the process" (p. 3). The test-by-test report style remains the most common technique used by psychologists (Pelco et al. 2009), despite being cited as problematic in the literature (Postal et al. 2018). Test-by-test reports can be a symptom of weak clinical reasoning because psychologists do not integrate other sources of information (e.g., observational data, background information) with the test scores in a meaningful way that will tell a story as to why the clients are struggling, along with the strengths that support them. Meyer et al. (2001) provided a clear explanation of the role of tests within an assessment, stating that "[T]ests do not think for themselves, nor do they directly communicate with patients. As in the case of a stethoscope, a blood pressure gauge, or an MRI scan, a psychological test is a dumb tool, and the worth of the tool cannot be separated from the sophistication of the clinician who draws inferences from it and then communicates with patients and professionals" (p. 153).

Clinical reasoning is more than interpreting test scores. Test scores should be connected to other information, including how clients attained their scores, error analysis, observation, and reports from selves and others. These additional data support a clear argument for how the conclusions were made. Assessment should also integrate client characteristics and functioning and the contextual aspects of the client's strengths and challenges, in order to inform interventions (Wright et al. 2022). Unfortunately, when information is segmented into individual sections, and test scores are reported in isolation, it is unclear to the reader why the client is experiencing difficulties, making it difficult to generate useful recommendations (Wright 2021).

The magnitude of this issue is highlighted in Dailor and Jacob's (2011) survey of 208 school psychologists. Of the respondents, 37% read a report within the past year that listed the student's test scores with no accompanying interpretation; 34% read reports that made recommendations that were unsubstantiated by the data, and 26% read computer-generated reports. Such reports are not useful to readers who depend on them to support clients through follow-up intervention. Limiting the reporting of findings to a list of strengths and weaknesses in the form of test scores reduces the role of the psychologist to that of a psychometrist (Wright et al. 2022). Instead, EBA should utilize an iterative hypothesis-testing and decision-making process that requires well-developed clinical reasoning skills (Suhr 2015; Wright et al. 2022).

4. How Do Psychologists Gain Clinical Reasoning Skills?

As a primarily invisible process, identifying how clinical reasoning skills develop through training and experience has been a challenge for both researchers and trainers. This might be the reason why programs spend more time assessing trainee proficiency in test administration than time assessing their broader assessment skills. In addition, there seems to be uncertainty about how or when trainees should learn clinical reasoning skills. Even though clinical reasoning is universally viewed as an important competence outcome by training programs (Harding 2007), programs do not necessarily have a systematic approach

to instruction. For instance, there is disagreement as to whether this should be taught in coursework or if it should be acquired through applied experiences such as practica and internship placements. The majority of clinics, schools, and neuropsychologists include assessment in their practice (Arocha and Patel 1995), yet a survey of clinical psychology programs found that less than half of the programs indicated that they teach strategies to improve decision-making and clinical judgment (Harding 2007). This is concerning because it is unlikely that clinical reasoning develops independently, without specific training (Harding 2007). Although the dominant view was once that students acquire these skills unconsciously via clinical experience (Wright 2021), there is growing recognition of the need to explicitly instruct and help trainees to develop accurate clinical reasoning.

Pre-doctoral internships also constitute an opportune period for developing clinical reasoning skills; pre-doctoral internships are generally a time to help students address areas of weakness, in order for them to enter the field with beginning levels of competence. Unfortunately, only 40% of APPIC internship sites offered intensive assessment training for interns (Krishnamurthy et al. 2004). Harding (2007) noted that this lack of training leads to significant concerns about practitioners' clinical reasoning because, without instruction in this area, psychologists are not likely to realize that they need to improve their clinical reasoning, and consequently, do not actively work to improve their clinical reasoning as they gain more experience. This poses a significant obstacle to psychologists' ability to provide EBA (Cook et al. 2017). As suggested by Gambrill (2012), clinicians are often unaware of the skills that they are lacking without specific feedback. Consequently, the current research suggests that psychologists do not generally receive enough training in clinical reasoning for assessment during their tenure in graduate programs to gain competence in this area.

4.1. Gaining and Measuring Clinical Reasoning

One of the issues with how clinical reasoning in assessment is taught (or not taught), is the limited understanding of what differentiates novices from experts and how much experience or what types of experiences are needed for someone to reach an "expert" level of practice. Researchers have struggled to effectively measure how reasoning develops from novice to expert. There has been an assumption that greater experience results in better clinical reasoning. Practitioners who have more experience should make fewer errors in reasoning and be able to identify what information is important and what legitimately contributes to the overall diagnostic picture. To examine this assumption, some researchers have focused on comparing the differences between experts and novices regarding diagnostic accuracy and reasoning processes.

4.2. Diagnostic Accuracy

In comparing the rates of diagnostic accuracy between less experienced clinicians and more experienced clinicians, the underlying assumption is that if the diagnosis is accurate, the clinical reasoning that proceeded it should be accurate as well. However, evaluating the accuracy of diagnostic decisions provides no information about *how* clinicians arrive at their conclusions (Siegert 1999). A focus on diagnostic accuracy is similar to an "outcome bias," which values outcomes over the quality of the process (Gambrill 2012, 2019). It relegates clinical reasoning to a "black box" where testing information enters and diagnostic conclusions exit, but the transformation process (e.g., clinical reasoning) is a mystery (Siegert 1999; Wright 2021).

Similar to the issues discussed earlier with the test-by-test report-writing style, this emphasis on outcome suggests a process that is directed by test scores, which results in minimizing or neglecting the role of the psychologist in taking responsibility for critically interpreting all of the data, not merely the test scores (Siegert 1999). The narrow focus on diagnostic accuracy fails to identify key differences and issues with the questions that psychologists choose to answer, the tools that they use, and the critical reasoning required

to make those decisions and integrate and interpret that information to describe client functioning and to make relevant recommendations.

4.3. The Role of Expertise in Clinical Reasoning

Without understanding the clinical reasoning required throughout the assessment process, it is difficult to identify which reasoning practices need to be targeted in training to improve diagnostic accuracy (Siegert 1999). In response, a small body of psychology research has studied the quality of clinical reasoning by examining the reasoning *processes* of practitioners. As with diagnostic accuracy, much of the literature has compared the processes of less experienced with more experienced practitioners. Within the broader literature, there are mixed findings regarding the effect of experience on the process of clinical reasoning.

A study of therapists found that expert therapists specializing in cognitive-behavioral and psychodynamic approaches generated more comprehensive and complex case conceptualizations than did both experienced therapists and trainees (Eells et al. 2005). A study by Arocha and Patel (1995) found that when trainees received contradictory information during case conceptualization, they were unsure how to manage it. Rather than adjusting their hypotheses, they tended to either ignore contradictory findings or interpret those findings to fit their initial hypothesis, rather than adjusting their hypothesis (Arocha and Patel 1995). Trainees also rigidly adhered to rules, paying little attention to contextual factors and, consequently, lacked discretionary judgment (Del Mar et al. 2006). Competent psychologists also demonstrated more skill in coping with pressures, having a broader conceptual framework for their planning, and following general standardized procedures.

The relatively sparse corpus of research focused specifically on psychoeducational assessment suggests that experience leads to limited improvements in clinical reasoning (de Mesquita 1992). For example, a study by Aspel et al. (1998) used a case-based approach to examine the process of clinical reasoning during psychoeducational assessment. Less and more experienced practitioners used similar approaches to the cases and did not change their working hypotheses after reviewing four to five categories of information. In another study, de Mesquita (1992) found experienced school psychologists, with varying levels of education, who considered similar types and amounts of information and came to similar conclusions as less experienced school psychologists. These two studies highlight the fact that experience does not automatically result in expertise. Education and experience were generally unrelated to diagnostic accuracy, and there was little difference among groups in terms of the amount and type of information reviewed and the number of diagnoses made.

However, when de Mesquita (1992) evaluated the process of clinical reasoning undertaken by practitioners, there were differences between less and more experienced practitioners. Practitioners with more experience required less time to reach an accurate diagnostic decision than did students. More experienced psychologists also generated fewer hypotheses and favored one hypothesis based on previous case experience. de Mesquita proposed that experience alone was not beneficial; instead, it was how well that knowledge was conceptually organized that led to accuracy and efficient reasoning.

Although experience seems to benefit psychologists in some ways, it is unclear how much experience is needed for someone to reach an expert level of practice, or if most practitioners even reach that level. Experience can support improvement, but it does not automatically lead to expertise. In medicine, Haynes et al. (2002) noted that expertise is not equivalent to experience. *Expertise* should be judged on one's knowledge of both the quality of the evidence and skill in interpreting that evidence, considering specific patient circumstances (Haynes et al. 2002). Tracey et al. (2014) found that practitioners gained confidence in their abilities along with experience, but their level of confidence did not match their performance. In fact, after gaining initial skills, confidence increased much more rapidly than accuracy, so the practitioners believed that they were more accurate than they actually were (Sanchez and Dunning 2018). Furthermore, confidence reduced their motivation to reflect on their skills, identify areas of weakness, and actively work

to improve them (Tracey et al. 2014). Without awareness of their limitations, clinicians were likely to continue to make the same mistakes after ten years of practice that they made in their first year because there was no opportunity for self-correction (Harding 2007; Watkins 2009). This highlights the importance of separating experience and expertise in understanding the role of clinical reasoning in EBA.

In summary, there is still much uncertainty about how experience and training influence the development of clinical reasoning as trainees move from graduate school to independent practice. The current literature suggests that the profession of psychology has approached clinical reasoning development in an ad hoc way. Relying on practical experiences (i.e., practica) for clinical reasoning development without intentional instruction or opportunities for feedback and reflection has the potential for ineffectual habits to become established, overconfidence to develop in practitioners, and little or no growth over time.

5. Moving Clinical Reasoning Skills from Novice to Expert

Research demonstrates that gaining expertise requires an intentional effort in learning and applying the component skills (Chow et al. 2015; Ericsson 2018; Miller et al. 2020) rather than acquiring clinical reasoning skills through supervised practice and then continued independent practice, which appears to the primary vehicle for learning clinical reasoning skills in psychology (Gross et al. 2019; Harding 2007; Krishnamurthy et al. 2004). Consequently, these findings suggest that to gain expertise in clinical reasoning, students require direct instruction and DP rather than simply additional experience. Unfortunately, there is currently no reliable model of assessment for clinical reasoning skills, which makes it difficult to determine where students or psychologists need to improve or how to help them to improve (Miller et al. 2020). As a result, the arguments presented in this section are largely based on research from other areas, and additional research is needed to identify how best these findings might apply to psychoeducational assessment.

Deliberate Practice

A body of research has examined the benefits of DP on expertise development in a variety of fields, including sports, performing arts, and chess (Ericsson 2018). DP requires clearly defining the individual components of the skill to be learned, immediate feedback in performing the skills, repeated practice of the skills, often in solitary settings, and using information from errors to improve performance (Ericsson 2006). In psychology, the outcomes of using DP in assessment have not yet been studied, although it has been successfully applied to psychotherapy practice. The amount of time that psychologists engaged in solitary DP (e.g., reviewing challenging cases, reviewing therapy recordings, writing down reflections and goals) predicted positive client outcomes during psychotherapy (Chow et al. 2015; Clements-Hickman and Reese 2020). It was more influential than other psychologist demographic variables, including experience, education, race, gender, and theoretical orientation. It is important to note that in DP, solitary practice is informed by feedback and coaching (Ericsson 2018; McLeod 2021; Miller et al. 2020). This was the only psychologist activity that predicted client outcomes and demonstrates both the importance of DP and the difference between experience and expertise.

The main components of DP are "(a) individualized learning objectives, (b) use of a coach, (c) feedback, and (d) successive refinement through repetition" (Miller et al. 2020, p. 39). Goal quality is related to performance levels, wherein the weakest performers do not generally engage in goal setting; average performers create goals focused on the desired outcome without setting smaller proximal goals; the highest performers set goals that break down the larger goal into steps that they will take to achieve the final outcome (Ericsson 2018). The research on implementing DP in therapy uses coaching with feedback because coaches are able to see aspects of performance that are often not evident to the psychologist. Beyond the typical requirements of feedback, such as specificity and timeliness, the feedback should focus on improving specific skills rather than on the final product, refining parts of the clinical reasoning process one step at a time, which leads to better performance in

the long run (Miller et al. 2020). One challenge with this process, especially for practicing psychologists, is that implementing changes will result in some failures due to the learning process. This requires a willingness to experience short-term failure in order to improve over the long term (Miller et al. 2020). Instead of focusing solely on how to assess, DP would direct attention to developing the psychologists' clinical reasoning (Miller et al. 2020). This process of DP has not yet been applied to assessment, but its success in therapy suggests that it is worth exploring this process in the context of assessment.

As with other practices, DP requires intentionality. Miller et al. (2020) offer suggestions for incorporating DP, including scheduling time for it, and protecting it by removing other distractions (e.g., emails or booking another meeting during that time). Taking time every week to jot down notes about what was learned through clinical practice, including successes as well as mistakes that were made and what contributed to them, is one example of an intention DP. Research is needed to determine how to effectively incorporate DP into clinical reasoning during assessment because it is an environment providing limited feedback (Lillienfeld and Basterfield 2020; Tracey et al. 2014). One strategy to improve the awareness of accuracy is to record and monitor one's diagnostic accuracy and utility over time (Kleinmuntz 1990); unfortunately, psychologists rarely receive this type of feedback from their psychological assessments (Mash and Hunsley 2005), and there is generally a low to moderate level of diagnostic agreement between clinicians (Rettew et al. 2009), making it exceedingly difficult for them to implement this strategy. More work is needed to find effective ways for psychologists to elicit feedback that they can use to inform their evaluations of their assessment practices.

One study found that explicitly teaching medical students how to engage in DP increased their planning and the structure of their work, as well as their performance on clinical exams (Duvivier et al. 2011). However, instruction was only as effective as the student's engagement with the process and required training in the self-assessment of weaknesses. Not surprisingly, students who were more accurate in their self-assessments performed better than students who were less accurate in their self-assessments (Duvivier et al. 2011).

6. Recommendations for Improving Clinical Reasoning

The first recommendation for improving clinical reasoning is to seek feedback throughout the assessment process and after the assessment is over. The nature of brief assessment relationships requires that psychologists intentionally and effortfully seek out this feedback (Siegert 1999). As noted in the work on DP in therapy, it is necessary to seek out negative feedback in order to identify areas of growth, which is necessary to improve practice (Miller et al. 2020). Mental health professionals often fail to acknowledge the uncertainty inherent in the assessment process (Gambrill 2012). Uncertainty throughout the process is inevitable because psychologists work under time constraints, using information of varying quality and completeness, but the negative impact of uncertainty is greater when psychologists fail to acknowledge that it exists (Gambrill 2012). As a result, professionals often overestimate their effectiveness, and those who are the most experienced are both the most confident and the least likely to be attentive to learning from their mistakes (Miller et al. 2020). In fact, overconfidence is one of the cognitive biases garnering the most research, making it an important area for psychologists to consider in their practice (Kahneman et al. 2021).

6.1. Framing the Assessment

From the outset, psychologists need to create the space and conditions for effective clinical reasoning. Of particular importance is the intentional practice to move away from the narrow framing of a case (e.g., "Does the client have _____ diagnosis?") because it similarly narrows the hypotheses generated, data collected, and the data that are considered (Gambrill 2012). Heath and Heath (2013) have argued that when individuals hold one hypothesis, all of their "ego" is invested in it, making it more challenging to actively attempt to disprove it or to pay attention to disconfirming information, increasing the likelihood of

engaging in confirmation bias. Putting forth a single hypothesis results in that hypothesis representing them as professionals, making it hard to be open to the possibility that their proposed hypothesis is incorrect. In contrast, developing multiple hypotheses allows the professionals' egos to be spread across the hypotheses, so as to allow their professional egos to be protected should one or more of their hypotheses be disconfirmed. In order to fully consider multiple hypotheses and to acknowledge the uncertainty inherent in assessment, it may be beneficial to ask what would need to be true for each of them to be the correct diagnosis, making sure to consider those hypotheses in which the psychologist does not initially have much confidence (Heath and Heath 2013).

Opening this space from the outset requires psychologists to reflect on their own assumptions about the client, referral question, and their goals versus client goals, in order to take steps to minimize bias and improve clinical reasoning (Gambrill 2019). It is important for psychologists to identify their assumptions about the client or about the presenting problems so that they can work to move beyond asking questions that reflect their beliefs rather than listening to the actual questions the client would like to have answered (Gambrill 2012). Consideration should also be given to noting potentially negative aspects of the process for clients, including the fact that accessing services may still be challenging after receiving a diagnosis and that recommendations generally require time and effort for the clients and their families (Heath and Heath 2013). This process requires strong listening skills and using motivational interviewing principles to better understand what the client wants to know and the changes to which they are committed in their lives (Suarez 2011). Motivational interviewing has the additional benefit that it can be used to increase client participation and their willingness to engage with later recommendations because it involves the psychologist taking the time to understand client goals and their willingness to make changes; it empowers clients to collaboratively engage in the assessment process (Suarez 2011).

6.2. Data Collection

Addressing cognitive biases in clinical practice is beyond the scope of this paper (see Gambrill 2012; 2019; Wilcox and Schroeder 2015). However, the most frequently noted strategy to improve clinical reasoning is to intentionally and systematically seek out information that could disprove the hypothesis, which relates to confirmation bias (Kleinmuntz 1990). Confirmation bias is a common contributor to making poor decisions because, when psychologists invest time and energy in pursuing a single hypothesis, they also invest their ego in it, which makes it more difficult to let the hypothesis go if there is disconfirming evidence. Humans are good at convincing themselves that they are collecting data in order to make a decision, when they are actually garnering support for the decision that they have already made (Heath and Heath 2013), making it important to take intentional steps to acknowledge and minimize confirmation bias in practice. Over-collecting data increases confidence without decreasing the objective uncertainty (Gambrill 2012).

Many assessment errors are the result of inattention and distraction during the test administration or the overconfidence that, with experience, psychologists can administer the test with less active engagement (e.g., reading test instructions verbatim; Oak et al. 2019). As noted above, acknowledging that all psychologists, including ourselves, are at risk of errors, rather than engaging in blind spot bias (e.g., "Others make errors, but I don't"), is the first step to the increasing awareness of errors and in taking steps to reduce them (Gambrill 2012). It is also important to remember that assessment is more than merely testing (Suhr 2015; Wright 2021). Assessment requires choosing measures to answer specific questions related to hypotheses from case conceptualization, actively approaching the data as a detective, attending not only to the psychometric properties of the measures but also attending to contextual and individual factors and the psychology of human behavior, which includes test scores as one source of data among many (Canivez 2019; Suhr 2015; Wright 2021).

6.3. Interpretation and Decision-Making

Psychologists face pressure to find answers for clients to support them in their difficulties, which can make psychologists feel as though they have to provide definitive answers. Psychologists, however, should beware of extremely high levels of confidence in predictive accuracy (Kleinmuntz 1990); they should, instead, practice humble acknowledgment of the limitations of the data available and of human judgment. In line with the ideals of Socratic ignorance, also known as Socratic wisdom, we should acknowledge the limits of the certainty of our conclusions because, as Popper (1996) noted, " . . . in our infinite ignorance, we are all equal" (p. 5). It is important to remember that there is always uncertainty during assessment; failing to acknowledge that uncertainty can increase errors (Gambrill 2012). We should also make sure to attend to contextual factors rather than only focusing on individual factors within the client, such as data from testing (Gambrill 2012). Finally, psychologists should consider documenting their decision-making process at each step, to increase transparency and access to information that could reveal errors, providing the opportunity to learn from them rather than repeat them (Kahneman et al. 2021). Psychologists should consider several questions to ensure that assessment findings are useful for clients, asking themselves: Do these findings and diagnoses help clients to better understand themselves? Do they inform recommendations that the clients are likely to follow? Do these findings make the clients and their families feel empowered (Nelson 2021)?

6.4. Considering Base Rates

Base rates represent one available tool to support clinical reasoning and increase diagnostic accuracy. Meehl (1957) argued that psychologists make more accurate decisions when they use base rates, rather than when they use clinical judgment. Consideration of "the relative frequency of phenomena" or of disorders and behaviors in a population (i.e., base rates; Kamphuis and Finn 2002) is important to consider because many psychologists work in clinical settings where almost all clients are presenting with a problem, making it easy to forget what is typical and what is abnormal in a population.

Base rate fallacy or base rate neglect occurs when practitioners do not use base rates when diagnosing; this results in false positives or negatives in the diagnostic decisions (Koehler 1996). Inattention to base rates is more likely to lead to poor decisions when the base rates conflict with other diagnostic information than when the data are in concordance. Koehler (1996) concluded that decision-makers are often accurate in situations with ample data and when these data are in line with base rates. They are, however, more prone to errors when the base rates are very different from their data. Base rate data can also be challenging due to the complexity of comorbidities that clients present with and the lack of operational definitions of the criteria for disorders (Ward 2019).

When base rate data are available, it is often aggregated (i.e., across the population). This provides the benefit of reducing the bias of individual clinics or psychologists (Reynolds 2016), but it may also obscure actual differences in base rates in a clinical setting as normative-based research sometimes hides individual differences, making them less useful for diagnostic purposes (Ward 2019). In order to effectively use base rates, psychologists need to have information that is specific to their type of practice. For example, the base rate of a specific disorder will be very different in a general practice than in a clinic specializing in a specific disorder, and there may be differences based on other demographic data (e.g., sex, geographical region, ethnicity, age (Youngstrom and Van Meter 2016)).

Although clinicians should consider base rates as part of EBA, there are some noted limitations. First, most studies looking at base rate neglect have been conducted in laboratory settings to find errors (Koehler 1996), leading to a limited understanding of the conditions under which base rate neglect occurs in real-life settings. A lack of information about the occurrence in practical settings makes it unclear how often base rate neglect is a problem, suggesting that the problem might be overemphasized in the research (Koehler 1996). Second, there are no clear guidelines or formulas that psychologists can use to apply base-rate

information in their practice (Kleinmuntz 1990). Third, during assessments, psychologists not only diagnose but provide information on the client's strengths and weaknesses, functioning, and prognosis, which cannot be accounted for by base rates (Garb and Schramke 1996). Further, research is needed to elucidate how to effectively incorporate base rates into practice.

6.5. Recommendations and Feedback

Building on the previous discussion of DP, psychologists should seek feedback throughout the assessment process and after the assessment is over. The brief nature of the assessment relationship requires that psychologists intentionally and effortfully seek out this feedback (Siegert 1999). As noted in the work on DP in therapy, it is necessary to seek out negative feedback in order to identify areas of growth, which is necessary to improve practice because psychologists are not likely to receive this important feedback as a matter of course (Miller et al. 2020).

Although not yet a common practice connected to psychoeducational assessments, there is a value in later connecting with clients to assist with the evaluation of clinical reasoning skills in relation to improved client functioning. To maximize the client's uptake of recommendations, one should be transparent in providing clients with evidence for the effectiveness of an assessment and recommendations, so that clients can make informed decisions (Gambrill 2012). Only 5% of clients think that psychologists' recommendations are helpful (Postal et al. 2018); when there are five recommendations, the clients will follow just over half of them (Elias et al. 2020). Even worse, about a third of clients do not follow any of the recommendations (Elias et al. 2020). Consequently, it is important to consider how psychologists can use clinical reasoning to improve the usability of recommendations. It may be helpful to work with clients to prioritize recommendations with clients and to engage in premortem planning to identify potential barriers, to ensure that they answer meaningful questions (Heath and Heath 2013), asking clients to think ahead, imagining that they did not implement the recommendation, and identifying what might prevent them from implementing the intervention. Then, the practitioner should work with the client to come up with solutions for each of those barriers. Conversely, it is also possible to ask clients to think ahead and pretend that they did implement the recommendation, and to identify what helped them to implement it. Then, we should work with clients to come up with ways to maximize those supports. This process complements motivational interviewing techniques by empowering clients to identify the recommendations that are the most meaningful to them, and encourages them to take an active role in determining the implementation of recommendations (Suarez 2011).

7. Conclusions

Clinical reasoning is an integral part of EBA that is currently poorly understood. As a result, there is little information on how psychologists develop clinical reasoning, how to assess the quality of clinical reasoning during an assessment, or how to gain and improve clinical reasoning skills. This has resulted in recommendations related to pieces of the assessment process, such as test administration, base rates, and report writing, without understanding the role of clinical reasoning in ensuring an EBA that supports clients. This paper outlines the current research in the area of clinical reasoning and draws from work in related fields to provide some initial suggestions on how to intentionally attend to clinical reasoning during an assessment. However, more work is needed to better understand the process of clinical reasoning in assessment, in order to determine the best ways to teach, monitor, and improve the clinical reasoning of psychologists during the assessment process.

Author Contributions: Conceptualization: G.W., M.S., M.A.D.; Writing Original Draft: G.W.; Writing Reviewing and Editing: G.W., M.S., M.A.D. 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 conflict of interest.

References

Ahmed, Adil H., Jyothsna Giri, Rahul Kashyap, Balwinder Singh, Yue Dong, Oguz Kilickaya, Patricia J. Erwin, M. Hassan Murad, and Brian W. Pickering. 2015. Outcome of adverse events and medical errors in the intensive care unit: A systematic review and meta-analysis. *American Journal of Medical Quality* 30: 23–30. [CrossRef] [PubMed]

American Psychological Association. 2006. Evidence-based practice in psychology. *American Psychologist* 61: 271–85. [CrossRef] [PubMed]

American Psychological Association. n.d. APA Dictionary of Psychology. Available online: https://dictionary.apa.org/critical-thinking (accessed on 10 June 2022).

Arocha, José F., and Vimla L. Patel. 1995. Novice diagnostic reasoning in medicine: Accounting for evidence. *The Journal of the Learning Sciences* 4: 355–84. Available online: http://www.jstor.org/stable/1466784 (accessed on 10 June 2022). [CrossRef]

Aspel, Andrew D., W. Grant Willis, and David Faust. 1998. School psychologists' diagnostic decision-making processes: Objective-subjective discrepancies. *Journal of School Psychology* 36: 137–49. [CrossRef]

Bornstein, Robert F. 2017. Evidence-based psychological assessment. *Journal of Personality Assessment* 99: 435–45. [CrossRef]

Burns, Candace W. 1990. Base rate theory and school psychology. *School Psychology Review* 19: 356–66. [CrossRef]

Canivez, Gary L. 2019. Evidence-based assessment for school psychology: Research, training, and clinical practice. *Contemporary School Psychology* 23: 194–200. [CrossRef]

Chow, Daryl L., Scott D. Miller, Jason A. Seidel, Robert T. Kane, and Jennifer A. Thornton. 2015. The role of deliberate practice in the development of highly effective psychotherapists. *Psychotherapy* 52: 337–45. [CrossRef]

Clements-Hickman, Alyssa L., and Robert J. Reese. 2020. Improving therapists' effectives: Can deliberate practice help? *Professional Psychology: Research and Practice* 51: 606–12. [CrossRef]

Cook, Jonathan R., Estee M. Hausman, Amanda Jensen-Doss, and Kristin M. Hawley. 2017. Assessment practices of child clinicians. *Assessment* 24: 210–21. [CrossRef]

Dailor, A. Nichole, and Susan Jacob. 2011. Ethically challenging situations reported by school psychologists: Implications for training. *Psychology in the Schools* 48: 619–31. [CrossRef]

de Mesquita, Paul D. 1992. Diagnostic problem solving of school psychologists: Scientific method or guesswork? *Journal of School Psychology* 30: 269–91. [CrossRef]

Del Mar, Chris, Jenny Doust, and Paul P. Glasziou. 2006. *Critical thinking: Evidence, communication, and decision-making*. Malden: Blackwell Publishing Inc.

Dombrowski, Stefan C., Ryan J. McGill, Ryan L. Farmer, John H. Kranzler, and Gary L. Canivez. 2021. Beyond the rhetoric of evidence-based assessment: A framework for critical thinking in clinical practice. *School Psychology Review*, 1–4. [CrossRef]

Duvivier, Robbert J., Jan van Dalen, Arno M. Muijtjens, Véronique R. M. P. Moulaert, Cees P. M. van der Vleuten, and Albert J. J. A. Scherpbier. 2011. The role of deliberate practice in the acquisition of clinical skills. *BMC Medical Education* 11: 101. [CrossRef]

Eells, Tracy D., Kenneth G. Lombart, Edward M. Kendjelic, L. Carolyn Turner, and Cynthia P. Lucas. 2005. The quality of psychotherapy case formulations: A comparison of expert, experienced, and novice cognitive-behavioral and psychodynamic therapists. *Journal of Consulting and Clinical Psychology* 73: 579–89. [CrossRef]

Elias, John, Eric Zimak, Andrea Sherwood, Beatriz MacDonald, Nubia Lozano, Jason Long, and A. Denise Larsen. 2020. Do parents implement pediatric neuropsychological report recommendations? *The Clinical Neuropsychologist* 35: 1117–33. [CrossRef]

Ericsson, K. Anders. 2006. The influence of experience and deliberate practice on the development of superior expert performance. In *The Cambridge Handbook of Expertise and Expert Performance*. Edited by K. Anders Ericsson, Robert R. Hoffman, Aaron Kozbelt and A. Mark Williams. Cambridge: University Press, pp. 685–705.

Ericsson, K. Anders. 2018. The differential influence of experience, practice, and deliberate practice on the development of superior individual performance of experts. In *The Cambridge Handbook of Expertise and Expert Performance*, 2nd ed. Edited by K. Anders Ericsson, Robert R. Hoffman, Aaron Kozbelt and A. Mark Williams. Cambridge: Cambridge University Press, pp. 745–69.

Gambrill, Eileen. 2012. *Critical Thinking in Clinical Practice: Improving the Quality of Judgments and Decisions*, 3rd ed. Hoboken: John Wiley and Sons.

Gambrill, Eileen. 2019. *Critical Thinking and the Process of Evidence-Based Practice*. New York: Oxford University Press.

Garb, Howard N., and Carol J. Schramke. 1996. Judgment research and neuropsychological assessment: A narrative review and meta-analyses. *Psychological Bulletin* 120: 140–53. [CrossRef]

Gordon, David, Joseph J. Rencic, Valerie J. Lang, Aliki Thomas, Meredith Young, and Steven J. Durning. 2022. Advancing the assessment of clinical reasoning across the health professions: Definitional and methodological recommendations. *Perspectives on Medical Education* 11: 108–14. [CrossRef]

Gross, Thomas J., Ryan L. Farmer, and Sarah E. Ochs. 2019. Evidence-based assessment: Best practices, customary practices, and recommendations for field-based assessment. *Contemporary School Psychology* 23: 304–26. [CrossRef]
Gruppen, Larry D. 2017. Clinical reasoning: Defining it, teaching it, assessing it, studying it. *The Western Journal of Emergency Medicine* 18: 4–7. [CrossRef]
Harding, Thomas P. 2007. Clinical decision-making: How prepared are we? *Training and Education in Professional Psychology* 1: 95–104. [CrossRef]
Haynes, R. Brian, P. J. Devereaux, and Gordon H. Guyatt. 2002. Clinical expertise in the era of evidence-based medicine and patient choice. *BMJ Evidence-Based Medicine* 7: 36–38. [CrossRef]
Heath, Chip, and Dan Heath. 2013. *Decisive: How to Make Better Choices in Lift and Work*. Toronto: Random House Canada.
Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgement*. New York: Little Brown Spark.
Kamphuis, Jan H., and Stephen E. Finn. 2002. Incorporating base rate information in daily clinical decision making. In *Clinical Personality Assessment: Practical Approaches*. Edited by James N. Butcher. New York: Oxford University Press, pp. 256–68.
Kleinmuntz, Benjamin. 1990. Why we still use our heads instead of formulas. *Psychological Bulleti* 107: 296–310. [CrossRef]
Koehler, Jonathan J. 1996. The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences* 19: 1–53. [CrossRef]
Krishnamurthy, Radhika, Leon Vande Creek, Nadine J. Kaslow, Yvette N. Tazeau, Marie L. Miville, Robert Kerns, Robert Stegman, Lisa Suzuki, and Sheryl A. Benton. 2004. Achieving competency in psychological assessment: Directions for education and training. *Journal of Clinical Psychology* 60: 725–39. [CrossRef]
Lillienfeld, Scott O., and Candice Basterfield. 2020. Reflective practice in clinical psychology: Reflections from basis psychological science. *Clinical Psychology: Science and Practic* 27: e12352. [CrossRef]
Mash, Eric J., and John Hunsley. 2005. Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Clinical Child and Adolescent Psycholog* 34: 362–39. [CrossRef]
McLeod, Julia. 2021. How students use deliberate practice during the first stage of counsellor training. *Counselling and Psychotherapy Research* 22: 1–12. [CrossRef]
Meehl, Paul E. 1957. When shall we use our heads instead of the formula? *Journal of Counseling Psychology* 4: 268–73. [CrossRef]
Meyer, Gregory J., Stephen E. Finn, Lorraine D. Eyde, Gary G. Kay, Kevin L. Moreland, Robert R. Dies, Elena J. Eisman, Tom W. Kubiszyn, and Geoffrey M. Reed. 2001. Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist* 56: 123–65. [CrossRef]
Miller, Scott D., Mark A. Hubble, and Daryl Chow. 2020. *Better Results: Using Deliberate Practice to Improve Therapeutic Effectiveness*. Washington, DC: American Psychological Association.
Nelson, Stephanie. 2021. *Advanced Report Writing [Webinar]*. Hopkinton: Massachusetts Neuropsychological Society. Available online: https://www.massneuropsych.org/content.aspx?page_id=22andclub_id=41215andmodule_id=448777 (accessed on 9 November 2021).
Oak, Erika, Kathleen D. Viezel, Ron Dumont, and John Willis. 2019. Wechsler administration and scoring errors made my graduate students and school psychologists. *Journal of Psychoeducational Assessment* 37: 679–91. [CrossRef]
Pelco, Lynn E., Sandra B. Ward, Lindsay Coleman, and Julie Young. 2009. Teacher ratings of three psychological report styles. *Training and Education in Professional Psychology* 3: 19–27. [CrossRef]
Popper, Karl. 1996. *In Search of a Better World: Lectures and Essays from Thirty Years*. New York: Routledge.
Postal, Karen, Clifton Chow, Sharon Jung, Kalen Erickson-Moreo, Flannery Geier, and Margaret Lanca. 2018. The stakeholders' project in neuropsychological report writing: A survey of neuropsychologists' and referral sources' views of neuropsychological reports. *The Clinical Neuropsychologist* 32: 326–44. [CrossRef]
Rettew, David C., Alicia Doyle Lynch, Thomas M. Achenbach, Levent Dumenci, and Masha Y. Ivanova. 2009. Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research* 18: 169–84. [CrossRef]
Reynolds, Cecil R. 2016. Contextualized evidence and empirically based testing and assessment. *Clinical Psychology: Science and Practice* 23: 410–16. [CrossRef]
Sanchez, Carmen, and David Dunning. 2018. Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology* 114: 10–28. [CrossRef]
Siegert, Richard J. 1999. Some thoughts about reasoning in clinical neuropsychology. *Behavior Change* 16: 37–48. [CrossRef]
Suarez, Mariann. 2011. Application of motivational interviewing to neuropsychology practice: A new frontier for evaluations and rehabilitation. In *The Little Black Book of Neuropsychology: A Syndrome-Based Approach*. Edited by Mike R. Schoenberg and James G. Scott. Boston: Springer, pp. 863–71.
Suhr, Julie A. 2015. *Psychological Assessment: A Problem-Solving Approach*. New York: Guilford.
Tracey, Terence J. G., Bruce E. Wampold, James W. Lichtenberg, and Rodney K. Goodyear. 2014. Expertise in psychotherapy: An elusive goal? *American Psychologist* 69: 218–29. [CrossRef]
Victor-Chmil, Joyce. 2013. Critical thinking versus clinical reasoning versus clinical judgment: Differential diagnosis. *Nurse Educator* 38: 34–36. [CrossRef]
Ward, Thomas J. 2019. EBA: Good idea but is it feasible? *Contemporary School Psychology* 23: 190–93. [CrossRef]

Watkins, Marley W. 2009. Errors in diagnostic decision making and clinical judgment. In *The Handbook of School Psychology*, 4th ed. Edited by Terry B. Gutkin and Cecil R. Reynolds. Hoboken: John Wiley and Sons Inc., pp. 210–29.

Wilcox, Gabrielle, and Meadow Schroeder. 2015. What comes before report writing? Attending to clinical reasoning and thinking errors in school psychology. *Journal of Psychoeducational Assessment* 33: 652–61. [CrossRef]

Wright, A. Jordan, Hadas Pade, Emily D. Gottfried, Paul A. Arbisi, David M. McCord, and Dustin B. Wygant. 2022. Evidence-based clinical psychological assessment (EBCPA): A review of the current state of the literature and best practices. *Professional Psychology: Research and Practice* 53: 372–86. [CrossRef]

Wright, A. Jordan. 2021. *Conducting Psychological Assessment: A Guide for Practitioners*, 2nd ed. Hoboken: Wiley.

Young, Meredith E., Aliki Thomas, Stuart Lubarsky, David Gordon, Larry D. Gruppen, Joseph Rencic, Tiffany Ballard, Eric Holmboe, Ana Da Silva, Temple Ratcliffe, and et al. 2020. Mapping clinical reasoning literature across the health professions: A scoping review. *BMC Medical Education* 20: 107. [CrossRef] [PubMed]

Youngstrom, Eric A., and Anna Van Meter. 2016. Empirically supported assessment of children and adolescents. *Clinical Psychology: Science and Practice* 23: 327–47. [CrossRef]

Youngstrom, Eric A., Sophia Choukas-Bradley, Casey D. Calhoun, and Amanda Jensen-Doss. 2015. Clinical guide to the evidence-based assessment approach to diagnosis and treatment. *Behavioral Practice* 22: 20–35. [CrossRef]

Journal of
Intelligence

MDPI

Article

Cognitive and Developmental Functions in Autistic and Non-Autistic Children and Adolescents: Evidence from the Intelligence and Development Scales–2

Salome D. Odermatt [1,*], Wenke Möhring [1,2], Silvia Grieder [1] and Alexander Grob [1]

[1] Department of Psychology, University of Basel, 4055 Basel, Switzerland
[2] Department of Educational Psychology and Health Psychology, University of Education Schwäbisch Gmünd, 73525 Schwäbisch Gmünd, Germany
* Correspondence: salome.odermatt@unibas.ch

Abstract: Autistic individuals often show impairments in cognitive and developmental domains beyond the core symptoms of lower social communication skills and restricted repetitive behaviors. Consequently, the assessment of cognitive and developmental functions constitutes an essential part of the diagnostic evaluation. Yet, evidence on differential validity from intelligence and developmental tests, which are commonly used with autistic individuals, varies widely. In the current study, we investigated the cognitive (i.e., intelligence, executive functions) and developmental (i.e., psychomotor skills, social–emotional skills, basic skills, motivation and attitude, participation during testing) functions of autistic and non-autistic children and adolescents using the Intelligence and Development Scales–2 (IDS-2). We compared 43 autistic (M_{age} = 12.30 years) with 43 non-autistic (M_{age} = 12.51 years) participants who were matched for age, sex, and maternal education. Autistic participants showed significantly lower mean values in psychomotor skills, language skills, and the evaluation of participation during testing of the developmental functions compared to the control sample. Our findings highlight that autistic individuals show impairments particularly in motor and language skills using the IDS-2, which therefore merit consideration in autism treatment in addition to the core symptoms and the individuals' intellectual functioning. Moreover, our findings indicate that particularly motor skills might be rather neglected in autism diagnosis and may be worthy of receiving more attention. Nonsignificant group differences in social–emotional skills could have been due to compensatory effects of average cognitive abilities in our autistic sample.

Keywords: autism spectrum disorder; cognitive functions; developmental functions; Intelligence and Development Scales–2; children and adolescents

Citation: Odermatt, Salome D., Wenke Möhring, Silvia Grieder, and Alexander Grob. 2022. Cognitive and Developmental Functions in Autistic and Non-Autistic Children and Adolescents: Evidence from the Intelligence and Development Scales–2. *Journal of Intelligence* 10: 112. https://doi.org/10.3390/jintelligence10040112

Received: 10 October 2022
Accepted: 16 November 2022
Published: 21 November 2022

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by difficulties in social communication and interaction accompanied by restricted repetitive behaviors, activities, and interests (American Psychiatric Association 2013). The worldwide prevalence of ASD has increased in recent years to approximately 1–2% (Idring et al. 2015; Maenner et al. 2020) and ASD is now considered a comparatively frequent condition (Happé and Frith 2020). Autistic individuals often experience difficulties beyond the core symptoms, such as impairments in cognitive and developmental domains, which in turn predict long-term development (e.g., Howlin and Moss 2012). Information about each individual's cognitive and developmental abilities is particularly important when it comes to making decisions about access to social services, the selection of appropriate treatment programs, and educational placement (White et al. 2007). Moreover, the amount of provided support is oftentimes determined on the basis of a cognitive assessment (Bowen 2014). According to the criteria of the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; American Psychiatric Association 2013) and the *International Statistical*

Classification of Diseases and Related Health Problems (11th ed.; World Health Organization 2018), clinicians have to report potential difficulties such as intellectual and language impairments in the diagnostic evaluation. Therefore, assessments with intelligence and developmental test batteries—in addition to autism-specific test procedures—represent a core part of the diagnostic process for autistic children and adolescents.

Yet, current tests for children and adolescents mainly allow the assessment of only single characteristics, such as intelligence, at a time and test batteries including multiple cognitive and developmental functions are missing so far. Consequently, when information about several domains or a broad assessment in a diagnostic evaluation is needed, clinicians often have to use various tests. This can be challenging, as the theoretical background and test administration differ widely among tests and dealing with these differences requires resources from the clinician. Moreover, tests build upon different characteristics of standardization samples and thus show less comparable scaled scores. The Intelligence and Development Scales–2 (IDS-2; Grob and Hagmann-von Arx 2018a) is a standardized test battery that assesses cognitive (i.e., intelligence and executive functions) and developmental (i.e., psychomotor skills, social–emotional skills, basic skills, motivation and attitude, and participation during testing) functions in 5- to 20-year-olds. The IDS-2 thus provides a comprehensive picture of an individual's strengths and difficulties with a single test battery across a wide age range from childhood to adolescence. In addition, the IDS-2 contains clear instructions and structured tasks, and many subtests use a closed-response format, which is particularly important for autistic children because of frequent structural language difficulties (Boucher 2012), making it suitable for administration with autistic individuals. Since the publication of the IDS-2 in 2018, it has often been used in psychological and medical practice in German-speaking countries. Further international adaptations for several other languages are currently in progress or have recently been published (e.g., Dutch, English, Italian, Polish; Grob et al. 2018, 2019, 2021, 2022). In the present study, we aimed to compare autistic children and adolescents to a matched non-autistic control sample on cognitive and developmental functions to study the differential validity of test scores from the IDS-2. By doing so, we can assess whether the IDS-2 is able to distinguish between clinical subgroups and typically developing individuals (Schmidt-Atzert and Amelang 2012).

Although general intellectual functioning varies substantially among autistic individuals, the latest report from the Centers for Disease Control and Prevention showed that almost 60% of autistic children are classified in the below-average intelligence range (IQ < 85), with about half of these children meeting criteria for intellectual disability (IQ $\leq$ 70; Maenner et al. 2020). Autistic individuals typically display uneven cognitive profiles, with relative strengths in nonverbal domains (e.g., Coolican et al. 2008; Grondhuis et al. 2018) and in tasks assessing abstract reasoning and visuospatial abilities (Charman et al. 2011; Nader et al. 2016), such as a well-documented peak in the Block Design subtest of the Wechsler Intelligence Scales (e.g., Muth et al. 2014). In contrast, relative weaknesses have been demonstrated in verbal domains, particularly in the Comprehension subtest of the Wechsler Intelligence Scales (e.g., Oliveras-Rentas et al. 2012), and in processing speed and working memory tasks[1] (Mayes and Calhoun 2003a; Nader et al. 2016; Oliveras-Rentas et al. 2012).

Autistic individuals often experience further cognitive difficulties on measures assessing executive functions (e.g., Hill 2004). Executive functions include a set of mental top-down regulation and control mechanisms (Miyake and Friedman 2012). In the theory of executive dysfunction, it is assumed that impairments in executive functions are responsible for some of the autism symptoms (Pennington and Ozonoff 1996), such as repetitive behavior (e.g., de Vries and Geurts 2012; Yerys et al. 2009). Demetriou et al. (2018) reported in the largest meta-analysis to date (235 studies) that autistic individuals showed moderate impairments in executive functions, both overall and in subdomains such as cognitive flexibility, fluency, planning, and inhibition,—which are also assessed with the IDS-2 (see Table S1 in the Supplement for an overview)—compared to non-autistic individuals.

Moreover, previous research showed significant impairments in autistic individuals' motor abilities, beginning in early childhood with deficits in the acquisition of motor milestones, such as later independent walking (e.g., Manicolo et al. 2019), and delays in gross and fine motor skills, for example, diminished object manipulation activity (Libertus et al. 2014; Provost et al. 2007). In a recent meta-analysis of 139 studies with samples of autistic children, adolescents, and young adults, their overall motor ability as well as gross and fine motor skills were strongly impaired in comparison to non-autistic peers (Coll et al. 2020). In line with this result, several studies found that autistic children, compared to non-autistic samples, scored lower on subscales (i.e., manual dexterity, ball skills, and balance) of the Movement Assessment Battery for Children–2 (M-ABC-2; Petermann 2008), which is a test of motor development that contains tasks similar to those in the IDS-2 psychomotor skills domain (Liu and Breslin 2013; Manicolo et al. 2019; Siaperas et al. 2012).

Further, research has shown that lower motor skills of autistic children were significantly associated with poorer social communication skills (MacDonald et al. 2013b). It has been suggested that motor problems might even precede social and communication deficits in autistic individuals because they may limit social participation and interaction with peers during play and may interfere with effective and timely movements, such as turning the head or pointing to something, that are particularly important for joint attention (Bhat et al. 2011). Impairments in social communication and interaction, such as difficulties in social–emotional reciprocity and nonverbal communicative behaviors, as well as in developing, maintaining, and understanding relationships constitute a core diagnostic characteristic of ASD (American Psychiatric Association 2013; World Health Organization 2018). These impairments are reflected in less accurate emotion recognition in human faces, with increased response times (Leung et al. 2022; Yeung 2022), more maladaptive emotion regulation strategies (Cai et al. 2018), including more reliance on others to regulate their emotions (Cibralic et al. 2019), and fewer socially competent behaviors (e.g., Meyer et al. 2009) compared to non-autistic individuals.

Additionally, language difficulties commonly co-occur with autism (Kjellmer et al. 2018). Some autistic individuals do not acquire verbal language at all (Brignell et al. 2018). Among those who develop language, delays often begin in infancy with retardations in the production of first words and in early language comprehension (e.g., Luyster et al. 2007; Mitchell et al. 2006). Moreover, across the preschool years, autistic children exhibit difficulties in phonological awareness skills (e.g., identifying syllables or onset-rimes), with slower development than their non-autistic peers (Dynia et al. 2019). Regarding language production and comprehension (i.e., expressive and receptive language skills, respectively), some studies indicated an atypical pattern, with better expressive and poorer receptive language skills in autistic individuals (e.g., Hudry et al. 2010). However, a meta-analysis examining 74 studies reported that autistic children and adolescents had scores that were approximately 1.5 standard deviations lower in receptive *as well as* expressive language abilities compared to non-autistic samples (Kwok et al. 2015).

In terms of academic skills, research indicated that autistic students demonstrate variable performance (Keen et al. 2016). Specifically, in previous studies, autistic individuals showed similar basic word-reading skills, such as word recognition, compared to non-autistic peers, but they tended to have difficulties in reading comprehension (for a meta-analysis: Brown et al. 2013). Autistic individuals with higher (vs. lower) reading skills also seemed to demonstrate better writing abilities (Zajic et al. 2020). Studies predominantly indicated deficits in text generation abilities for autistic individuals, while overall intact or slightly impaired spelling skills were reported (Finnegan and Accardo 2018; Mayes and Calhoun 2003a, 2003b). Similarly, the majority of autistic individuals exhibited average competencies in mathematics, such as mathematical problem solving, compared to non-autistic peers or to the norm population in previous research (Chiang and Lin 2007; Titeca et al. 2017; Troyb et al. 2014).

Concerning motivation and attitude, a recent meta-analysis reported that autistic individuals displayed significantly lower levels of conscientiousness than non-autistic

individuals (Lodi-Smith et al. 2019). In contrast, less is known regarding achievement motivation in autistic individuals. A few studies reported that autistic individuals encountered problems with self-regulation (e.g., Jahromi et al. 2012; Konstantareas and Stewart 2006) and displayed higher interest in mathematics while simultaneously showing more fear of failure and lower mastery goals (Georgiou et al. 2018). Moreover, autistic children tended to exhibit impaired engagement (Keen 2009), especially in assessment situations where they frequently demonstrated off-task behaviors (Akshoomoff 2006) and a lack of willingness to complete tasks (Mandelbaum et al. 2006).

Previous research has rarely used the IDS-2 in order to test autistic individuals. The only study so far reported in the technical manual of the IDS-2 (Grob and Hagmann-von Arx 2018b) built upon a small sample of autistic children and adolescents (N = 18; M_{age} = 13 years 4 months, age range 8–17 years; 17 males and 1 female). Findings showed significantly lower group mean values for autistic children and adolescents compared to non-autistic peers in the composite score of social–emotional skills (d = 0.62) and the composite score of psychomotor skills (d = 1.01) of the IDS-2. No differences were found in the composite scores of other domains. However, evidence of possible differences at the level of subtests is currently lacking, as analyses on this level have not been performed. Moreover, the study included mainly children and adolescents with Asperger's syndrome (n = 13) and no participants with previously diagnosed infantile autism. Given the small sample size, which may have diminished the power to find group differences, and the biased distribution of sex and subtype, it remains unknown to what extent these results can be generalized.

Building on this theoretical background, we pursued two goals for the present study: First, we aimed to extend previous research on various cognitive and developmental functions in autistic children and adolescents using a single test procedure and based on the norms of a large and representative standardization sample. By doing so, our findings will provide a comparable and comprehensive view of participants' performance in relevant domains. Second, we aimed to add knowledge regarding the differential validity evidence for test scores of the IDS-2 in autistic individuals, as psychological test procedures need to be examined in terms of their scientific quality in order to draw appropriate conclusions based on their test results. Given that previous research had some limitations (Grob and Hagmann-von Arx 2018b), we attempted to overcome these shortcomings by assessing a larger sample, including a more representative mapping of sex and subtypes, and performing analyses at the level of subtests, which have not yet been investigated in this population. We therefore examined possible mean-level differences between a large sample of autistic children and adolescents and a control sample of non-autistic children and adolescents matched by age, sex, and maternal education in the cognitive and developmental functions measured by the IDS-2. We included maternal education as a proxy for socioeconomic status (SES) to control for the fact that more autistic children and adolescents come from families with higher SES than from other SES groups (Thomas et al. 2012; Van Meter et al. 2010).

Taking into consideration the presented literature, we hypothesized that autistic children and adolescents would score lower than the control sample of non-autistic children in the following IDS-2 domains as displayed in Table 1, while we assumed that autistic children and adolescents' scores would be similar to those of the control sample in the other IDS-2 domains (see Table 1 for a summary).

Table 1. Summary of our hypotheses.

Domain	Assumed Differences in Performance between Autistic and Non-Autistic Participants	Assumed Similar Performance inAutistic and Non-Autistic Participants
	Variable	Variable
Intelligence	Composite scores (Profile IQ, Full-Scale IQ, Screening IQ) Processing Speed Parrots Boxes Auditory Short-Term Memory Digit and Letter Span Mixed Digit and Letter Span Visuospatial Short-Term Memory Shape Memory Rotated Shape Memory Verbal Reasoning Naming Categories Naming Opposites Long-Term Memory Story Recall Picture Recall	Visual Processing Shape Design Washer Design Abstract Reasoning Matrices: Completion Matrices: Odd One Out
Executive functions	Composite score Listing Words Divided Attention Animal Colors Drawing Routes	
Psychomotor skills	Composite score Gross Motor Skills Fine Motor Skills Visuomotor Skills	
Social–emotional skills	Composite score Identifying Emotions Regulating Emotions Socially Competent Behavior	
Basic skills	Language skills Phoneme Analysis Phoneme–Grapheme Correspondence Language Expressive Language Receptive Text Comprehension	Composite score Logical–Mathematical Reasoning Reading Reading Words Reading Pseudo Words Spelling
Motivation and attitude	Composite score Conscientiousness Achievement Motivation	
Participation during testing	intelligence executive functions developmental functions	

Note. Differences in performance between autistic and non-autistic participants are interpreted as meaningful if the *p* value is significant after Hommel's correction and the effect size is at least small.

2. Materials and Methods

2.1. Participants and Procedure

Forty-three autistic children and adolescents (M_{age} = 12 years 4 months, age range 7–17 years; 35 males and 8 females) were recruited during (n = 18) or after (n = 25) the IDS-2 standardization and validation study with the help of local child and adolescent psychiatric services and hospitals, privately practicing psychiatrists and psychotherapists who are

experts in autism diagnoses, and associations for autistic individuals. All included children and adolescents were diagnosed with ASD (infantile autism: n = 11, atypical autism: n = 6, Asperger's syndrome: n = 24, not specified: n = 2) but were not selected on the basis of specific subtypes. Participants had received the diagnosis on average 4.08 years (SD = 2.61) prior to their participation in the present study. The ratio of males to females corresponded to the distribution of approximately four males to one female diagnosed with ASD in the population (Maenner et al. 2020).

A control sample of 43 non-autistic children and adolescents (M_{age} = 12 years 6 months, age range 6–20 years; 35 males and 8 females) was drawn from the German standardization and validation sample of the IDS-2 (N = 2030; Mage = 12 years 3 months, age range 5–20 years; 977 males and 1053 females). The control sample was matched by age, sex, and maternal education (as a proxy for SES) and did not differ regarding demographic characteristics from the sample of autistic children and adolescents (see Table 2). Non-autistic children and adolescents were recruited from kindergartens and schools.

Table 2. Demographic Characteristics of Autistic and Non-Autistic Children and Adolescents.

Characteristic	Autistic Sample n = 43		Non-Autistic Sample n = 43		χ^2	p
	n	%	n	%		
Sex					11.33	1.000
Female	8	19	8	19		
Male	35	81	35	81		
Maternal education					14.24	1.000
No postsecondary education	23	54	23	54		
Compulsory school	1	2	2	5		
Apprenticeship	16	37	15	35		
High school	1	2	1	2		
Higher vocational education	5	12	5	12		
Postsecondary education (university degree)	19	44	19	44		
Other	0	0	0	0		
Unknown	1	2	1	2		
Participants' current education					7.00	1.000
Kindergarten	0	0	1	2		
Elementary school	14	33	20	47		
Secondary school	10	23	11	26		
School for special education	8	19	1	2		
High school	6	14	5	12		
Apprenticeship	3	7	4	9		
University	0	0	1	2		
None	2	5	0	0		
Intelligence level					11.09	1.000
<70	9	21	1	2		
70–84	7	16	6	14		
85–99	8	19	16	37		
100–114	11	26	14	33		
≥115	6	14	6	14		
Comorbid condition					12.15	1.000
Visual impairment	6	14	8	19		
Motor problems	4	9	0	0		
Speech problems	4	9	1	2		
Dyslexia	2	5	2	5		
Dyscalculia	0	0	2	5		
AD(H)D	4	9	4	9		
Depression	1	2	1	2		
Medical problems	10	23	2	5		
Ethnicity					10.04	1.000
German-speaking country	38	88	39	91		
Other European country	4	9	2	5		
Non-European country	1	2	1	2		
Unknown	0	0	1	2		

Table 2. *Cont.*

Characteristic	Autistic Sample n = 43		Non-Autistic Sample n = 43		χ^2	p
	n	%	n	%		
Native language					10.99	1.000
Monolingual German	32	74	35	81		
Bilingual	6	14	6	14		
Other language than German	5	12	2	5		

Note. Samples were matched for age, sex, and maternal education (as a proxy for socioeconomic status). Autistic sample: M_{age} = 12.3 years, SD = 3.08; non-autistic sample: M_{age} = 12.51 years, SD = 3.56. Paired-sample t test for age: t = 0.34, p = .733. χ^2 test for sex (0 = male, 1 = female), maternal education (0 = no postsecondary education, 1 = postsecondary education), participants' current education (0 = no special education, 1 = special education), intelligence level (0 = average, 1 = below/above average), comorbid condition (0 = no, 1 = yes), ethnicity (0 = German-speaking country, 1 = other), and native language (0 = monolingual, 1 = not monolingual). AD(H)D = attention deficit/hyperactivity disorder or attention deficit disorder.

All participants were individually tested using the IDS-2 by psychologists or trained psychology students. For the administration of the IDS-2 with autistic children and adolescents, we received input from psychiatrists and psychotherapists who specialize in autism. Test administration lasted approximately 4 h and was split into two sessions no longer than 1 week apart upon a participant's request. Participants were tested either at their homes or in a laboratory at the university. The local ethics committee (Ethics Committee Northwest and Central Switzerland) provided approval and the study was conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from participants and/or their parents.

2.2. Instrument

A detailed description of the IDS-2 (Grob and Hagmann-von Arx 2018a) can be found in the Supplemental Material (Table S1). Psychometric properties have been demonstrated in several studies for the standardization sample (Grieder and Grob 2020; Grob and Hagmann-von Arx 2018b). Demographic characteristics were assessed through a parental interview at the beginning of the first test session.

2.3. Statistical Analyses

Analyses were conducted with R (R Core Team 2021). To obtain a non-autistic sample that would be comparable to the autistic sample with respect to demographic characteristics, we performed a matching procedure using the MatchIt package (Ho et al. 2011). We matched the two samples by age (nearest; continuous), sex (exact; 0 = male, 1 = female), and maternal education (nearest; 1 = compulsory school, 2 = apprenticeship, 3 = high school, 4 = higher vocational education, 5 = university degree, 6 = other, 7 = unknown). We calculated independent-samples t tests to investigate mean-level differences between the autistic sample and the non-autistic sample in cognitive and developmental domains using standardized scores (M = 100, SD = 15, for Profile IQ, Full-Scale IQ, Screening IQ, and the seven intelligence group factors; M = 10, SD = 3, for other composite scores and subtests). To reduce the alpha error inflation caused by multiple testing, p values were adjusted with Hommel's (1988) correction by including p values from all tests simultaneously. Effect sizes were computed (Cohen 1988) and interpreted in accordance with common practice (Cohen's d; small effect: $d \geq 0.20$, medium effect: $d \geq 0.50$, large effect: $d \geq 0.80$). A post hoc power analysis using G*Power (Faul et al. 2007) revealed that with α = .05 and power = .80, small effects (d = 0.30) could be detected in the present sample (note that this is without accounting for multiple testing). Differences were interpreted as meaningful if they were significant after Hommel's correction and showed at least a small effect size. In addition, we reported reliabilities for all IDS-2 scores, consisting of Cronbach's alpha for homogeneous subtests; reliabilities calculated according to a formula of Lienert and Raatz (1998) for composite scores, which are based on intercorrelations and reliabilities of those subtests or tasks that are included in the corresponding score; or retest reliabilities reported in the

technical manual of the IDS-2 (Grob and Hagmann-von Arx 2018b) for subtests that contain a single score or consist of heterogeneous tasks.

3. Results

Reliabilities, descriptive statistics, and results of the independent-samples t tests[2] are presented in Table 3 for the cognitive functions and in Table 4 for the developmental functions. Reliabilities were high for composite scores and high-to-satisfactory for subtests in both samples.

Table 3. Reliabilities, Means, Standard Deviations, and t tests of the Cognitive Functions From the Intelligence and Development Scales–2 for Autistic and Non-Autistic Children and Adolescents.

Variable	Autistic Sample N = 43				Non-Autistic Sample n = 43				t	df	p	p_H	d
	Rel	M	SD	Range	Rel	M	SD	Range					
Profile IQ [b]	.99	90.16	19.98	55–131	.99	97.68	13.82	61–121	1.96	77	.027	.406	0.44
Full-Scale IQ [b]	.99	91.58	20.77	55–129	.98	97.63	13.60	63–120	1.58	81	.059	.627	0.35
Screening IQ[b]	.98	93.54	19.35	55–125	.98	100.58	16.45	61–134	1.80	82	.038	.490	0.39
Visual Processing [b]	.99	97.03	21.50	55–129	.97	102.67	12.50	80–129	1.46	80	.148	.809	0.32
Processing Speed [b]	.98	95.58	20.47	55–143	.98	100.19	15.57	56–126	1.15	80	.126	.758	0.25
Auditory Short-Term Memory [b]	.97	90.77	16.95	55–139	.96	97.76	12.54	64–121	2.12	79	.019	.352	0.47
Visuospatial Short-Term Memory [b]	.97	88.92	15.12	55–118	.94	96.79	10.89	77–118	2.70	79	.004	.161	0.60
Abstract Reasoning [b]	.98	95.10	20.67	55–141	.97	97.55	14.93	63–122	0.62	80	.539	.846	0.14
Verbal Reasoning [b]	.99	94.32	19.29	58–126	.97	99.98	15.29	61–131	1.48	81	.071	.674	0.33
Long-Term Memory [b]	.97	88.08	15.82	55–113	.97	93.64	16.06	58–137	1.57	79	.060	.627	0.35
Shape Design [a]	.95	9.62	4.24	1–16	.89	10.65	2.55	7–16	1.36	83	.176	.846	0.30
Washer Design [a]	.94	9.54	3.83	1–17	.92	10.57	2.78	4–19	1.41	81	.162	.811	0.31
Parrots [a]	.92	9.15	4.28	1–19	.91	9.74	2.94	1–17	0.75	82	.228	.846	0.16
Boxes [a]	.93	9.28	3.29	1–16	.90	10.40	3.36	2–17	1.54	80	.064	.642	0.34
Digit and Letter Span [a]	.90	9.02	3.67	1–18	.82	9.95	2.17	5–14	1.42	83	.079	.693	0.31
Mixed Digit and Letter Span [a]	.86	8.49	3.19	1–18	.84	10.21	2.99	1–17	2.51	79	.007	.231	0.56
Shape Memory [a]	.88	8.41	2.99	1–14	.78	9.28	2.36	5–16	1.47	82	.072	.674	0.32
Rotated Shape Memory [a]	.90	8.36	3.06	2–17	.82	10.12	2.63	6–18	2.78	79	.003	.132	0.62
Matrices: Completion [a]	.93	9.32	3.63	3–18	.90	10.70	3.07	4–17	1.89	82	.063	.628	0.41
Matrices: Odd One Out [a]	.93	9.72	3.82	2–18	.86	9.14	2.93	2–15	−0.78	80	.440	.846	0.17
Naming Categories [a]	.95	9.38	3.92	1–16	.92	10.33	3.49	2–18	1.17	83	.122	.755	0.25
Naming Opposites [a]	.92	9.41	3.49	1–16	.87	10.50	2.90	3–19	1.54	81	.063	.633	0.34
Story Recall [a]	.93	8.40	3.56	1–14	.88	9.42	3.17	1–16	1.39	82	.084	.693	0.30
Picture Recall [a]	.85	8.05	2.73	3–14	.88	8.90	3.32	3–18	1.27	80	.104	.726	0.28
Executive functions composite score [b]	.97	8.84	2.20	4–13	.96	9.95	1.98	6–15	2.27	71	.013	.317	0.53
Listing Words [c]	.75	7.89	3.13	1–14	.75	9.55	2.93	4–17	2.38	73	.010	.292	0.55
Divided Attention [b]	.92	8.76	2.93	4–15	.90	10.11	2.52	5–17	2.13	71	.019	.352	0.50
Animal Colors [c]	.72	8.09	3.55	1–14	.72	9.47	3.44	3–19	1.70	72	.047	.547	0.40
Drawing Routes [b]	.96	9.91	2.60	5–15	.94	10.41	2.41	5–15	0.88	74	.192	.846	0.20

Note. Samples were matched for age, sex, and maternal education (as a proxy for socioeconomic status). p_H indicates p values adjusted with Hommel's (1988) correction. Please note that after this correction, none of the comparisons were significant. Rel indicate reliabilities. The following reliabilities are reported: [a] Cronbach's alpha, [b] reliability calculated according to a formula by Lienert and Raatz (1998), or [c] retest reliability.

Table 4. Reliabilities, Means, Standard Deviations, and t tests of the Developmental Functions From the Intelligence and Development Scales–2 for Autistic and Non-Autistic Children and Adolescents.

Variable	Autistic Sample n = 43				Non-Autistic Sample n = 43				t	df	p	p_H	d
	Rel	M	SD	Range	Rel	M	SD	Range					
Psychomotor skills composite score [b]	.98	8.49	2.22	4–12	.95	10.43	1.57	7–15	4.60	81	<.001	**<.001**	**1.01**
Gross Motor Skills [a]	.72	5.29	3.57	1–11	.77	11.35	3.08	5–15	5.30	32	<.001	**<.001**	**1.82**
Fine Motor Skills [b]	.96	8.65	3.03	2–14	.96	10.59	2.39	4–16	3.20	79	<.001	**.046**	**0.71**
Visuomotor Skills [b]	.95	8.79	1.97	4–13	.87	10.01	1.73	7–13	3.01	81	.002	.077	0.66

Table 4. *Cont.*

Variable	Autistic Sample n = 43				Non-Autistic Sample n = 43				t	df	p	p_H	d
	Rel	M	SD	Range	Rel	M	SD	Range					
Social–emotional skills composite score [b]	.96	8.32	2.86	1–13	.95	9.82	2.15	5–13	2.71	82	.004	.154	0.59
Identifying Emotions [c]	.85	7.71	4.27	1–12	.85	10.24	2.68	4–12	2.07	32	.023	.388	0.71
Regulating Emotions [c]	.78	8.40	3.39	1–13	.78	10.00	2.75	4–15	2.37	82	.010	.292	0.52
Socially Competent Behavior [c]	.71	8.32	3.17	1–15	.71	9.81	2.70	5–15	2.29	80	.012	.311	0.51
Basic skills composite score [b]	.99	9.77	2.53	2–14	.99	9.92	2.15	5–13	0.29	74	.772	.846	0.07
Logical–Mathematical Reasoning [a]	.98	9.07	4.21	1–17	.96	10.24	3.07	3–16	1.44	81	.153	.809	0.32
Language Skills [b]	.98	7.12	2.65	3–14	.98	10.50	1.86	6–13	4.11	28	<.001	**.008**	**1.51**
Phoneme Analysis [a]	.92	6.07	3.00	1–13	.97	10.24	3.13	3–15	3.75	29	<.001	**.019**	**1.35**
Phoneme–Grapheme Correspondence [a]	.86	9.07	4.20	1–15	.96	10.00	2.18	5–13	0.79	29	.217	.846	0.29
Language Expressive [a]	.81	7.08	2.96	1–14	.89	10.41	2.55	5–15	3.31	28	.001	.058	1.22
Language Receptive [a]	.85	6.71	3.34	2–14	.81	11.35	2.37	7–16	4.52	29	<.001	**.003**	**1.63**
Reading [b]	.98	8.76	3.33	1–15	.95	9.65	2.40	5–14	1.35	74	.182	.846	0.31
Reading Words [c]	.79	9.34	3.41	2–16	.79	9.47	2.56	5–14	0.20	76	.846	.846	0.04
Reading Pseudo Words [c]	.67	8.89	2.89	2–14	.67	9.82	2.48	4–14	1.52	75	.132	.792	0.35
Text Comprehension [a]	.69	9.40	5.02	1–16	.69	10.37	2.79	4–16	1.00	68	.160	.809	0.24
Spelling [a]	.88	8.89	3.19	3–15	.88	9.79	2.66	4–15	1.26	65	.212	.846	0.31
Motivation and attitude composite score [b]	.96	10.56	3.24	6–17	.96	10.65	2.78	6–19	0.11	46	.458	.846	0.03
Conscientiousness [a]	.82	10.21	3.27	6–18	.79	10.26	2.85	6–19	0.06	45	.477	.846	0.02
Achievement Motivation [a]	.87	11.12	3.96	4–19	.86	11.04	3.11	6–19	−0.07	47	.528	.846	0.02
Participation during testing, intelligence [a]	.93	8.19	3.15	1–16	.93	10.17	3.57	1–16	2.68	81	.004	.169	0.59
Participation during testing, executive functions [a]	.89	8.76	2.75	1–16	.91	10.17	2.89	4–16	2.13	71	.018	.351	0.50
Participation during testing, developmental functions [a]	.95	8.33	3.31	1–16	.92	10.66	3.02	5–16	3.30	79	<.001	**.035**	**0.73**

Note. Samples were matched for age, sex, and maternal education (as a proxy for socioeconomic status). p_H indicates p values adjusted with Hommel's (1988) correction. Significant results after accounting for multiple testing (Hommel correction) are presented in bold. Rel indicate reliabilities. The following reliabilities are reported: [a] Cronbach's alpha, [b] reliability calculated according to a formula by Lienert and Raatz (1998), or [c] retest reliability.

3.1. Cognitive Functions

Figure 1 displays the means and standard deviations in the cognitive functions of the IDS-2 for the autistic and non-autistic samples. Before controlling for multiple testing, we found significant group differences for the intelligence composite scores: Profile IQ, $t(77) = 1.96$, $p = .027$, and Screening IQ, $t(82) = 1.80$, $p = .038$, with small effect sizes ($d = 0.44$ and 0.39, respectively), indicating lower scores for the autistic sample than the control sample. Furthermore, we observed group differences for the intelligence group factors: Auditory Short-Term Memory, $t(79) = 2.12$, $p = .019$, and Visuospatial Short-Term Memory, $t(79) = 2.70$, $p = .004$, with small-to-medium effect sizes ($d = 0.47$ and 0.60, respectively), and the corresponding subtests Mixed Digit and Letter Span, $t(79) = 2.51$, $p = .007$, and Rotated Shape Memory, $t(79) = 2.78$, $p = .003$, with medium effect sizes ($d = 0.56$ and 0.62, respectively), such that the autistic participants showed lower mean values than the control sample. Moreover, the autistic participants had significantly lower mean values in the executive functions composite score, $t(71) = 2.27$, $p = .013$, and the subtests Listing Words, $t(73) = 2.38$, $p = .010$, Divided Attention, $t(71) = 2.13$, $p = .019$, and Animal Colors, $t(72) = 1.70$, $p = .047$. Effect sizes were in the small-to-medium range ($d = 0.40$ to 0.55). We found no differences between autistic and non-autistic participants in the Full-Scale IQ, $t(81) = 1.58$, $p = .059$, in the intelligence group factors Visual Processing, $t(80) = 1.46$, $p = .148$, Processing Speed, $t(80) = 1.15$, $p = .126$, Abstract Reasoning, $t(80) = 0.62$, $p = .539$, Verbal Reasoning, $t(81) = 1.48$, $p = .071$, and Long-Term Memory, $t(79) = 1.57$, $p = .060$, including corresponding intelligence subtests, and in the executive functions subtest Drawing Routes, $t(74) = 0.88$, $p = .192$.

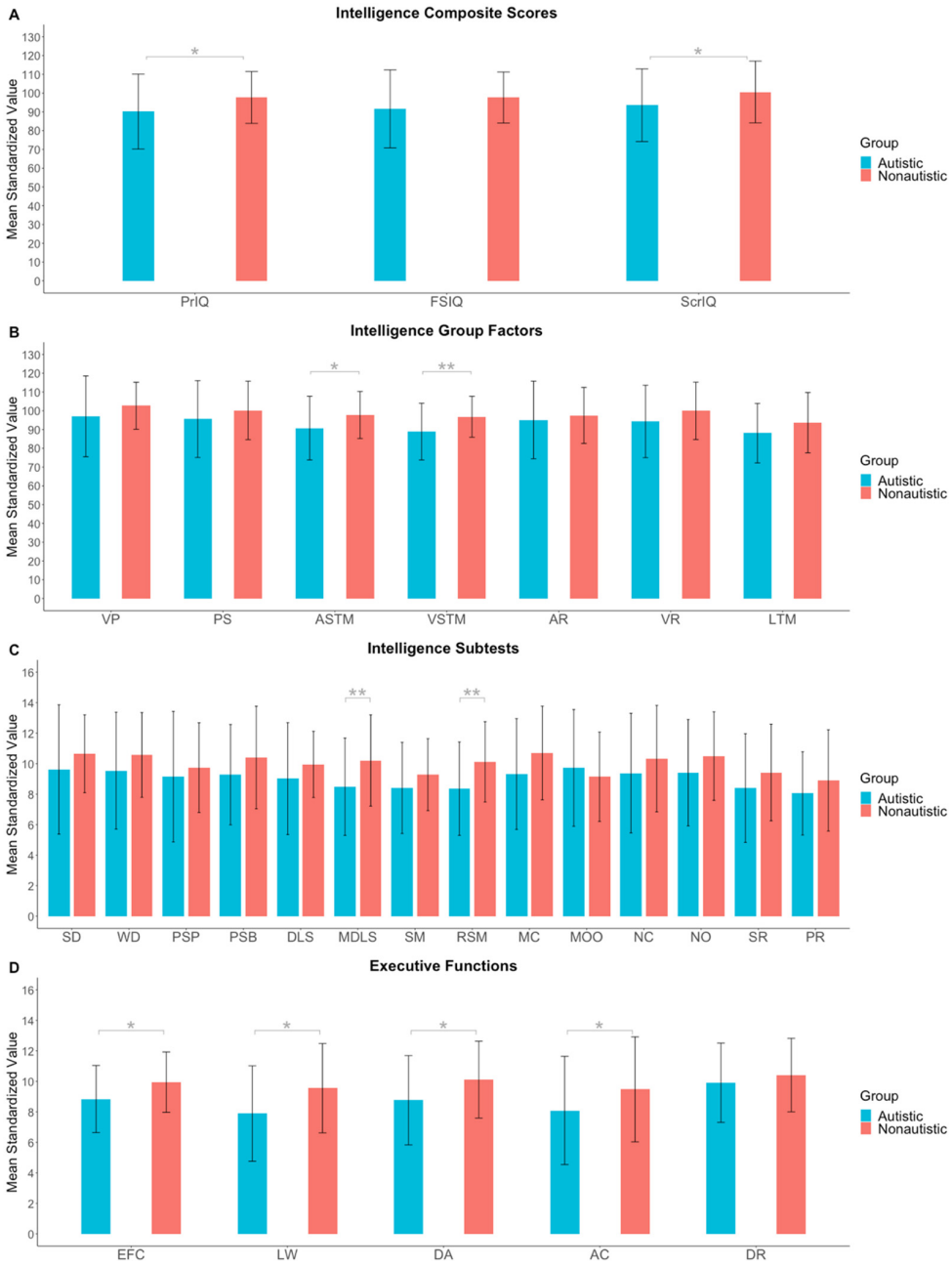

Figure 1. Means and standard deviations are reported for (**A**) intelligence composite scores, (**B**) intelligence group factors, (**C**) intelligence subtests, and (**D**) executive functions composite score and subtests of the Intelligence and Development Scales–2 for autistic and non-autistic children and adolescents. Asterisks in grey indicate p values not adjusted with Hommel's (1988) correction. Asterisks in black indicate p values adjusted according to Hommel (1988). Please note that after this correction, none of the comparisons were significant and therefore, no black asterisks are included in the present graphs. PrIQ = Profile IQ; FSIQ = Full-Scale IQ; ScrIQ = Screening IQ; VP = Visual Processing; PS = Processing Speed; ASTM = Auditory Short-Term Memory; VSTM = Visuospatial Short-Term Memory; AR = Abstract Reasoning; VR = Verbal Reasoning; LTM = Long-Term Memory; SD = Shape Design; WD = Washer Design; PSP = Parrots; PSB = Boxes; DLS = Digit and Letter Span; MDLS = Mixed Digit and Letter Span; SM = Shape Memory; RSM = Rotated Shape Memory; MC = Matrices: Completion; MOO = Matrices: Odd One Out; NC = Naming Categories; NO = Naming Opposites; SR = Story Recall; PR = Picture Recall; EFC = Executive functions composite score; LW = Listing Words; DA = Divided Attention; AC = Animal Colors; DR = Drawing Routes. * $p < .05$. ** $p < .01$. *** $p < .001$.

However, after controlling for multiple testing, the significant differences in intelligence and executive functions fell above the Hommel-corrected p-value threshold (see Table 3).

3.2. Developmental Functions

Figure 2 shows the means and standard deviations in the developmental functions of the IDS-2 for the autistic and non-autistic samples. Before controlling for multiple testing, results indicate that autistic participants scored significantly lower than non-autistic participants in psychomotor skills [composite score, $t(81) = 4.60$, $p < .001$; Gross Motor Skills, $t(32) = 5.30$, $p < .001$; Fine Motor Skills, $t(79) = 3.20$, $p < .001$; Visuomotor Skills, $t(81) = 3.01$, $p = .002$] with medium-to-large effect sizes ($d = 0.66$ to 1.82). We found a similar group difference for participants' social–emotional skills [composite score, $t(82) = 2.71$, $p = .004$; Identifying Emotions, $t(32) = 2.07$, $p = .023$; Regulating Emotions, $t(82) = 2.37$, $p = .010$; Socially Competent Behavior, $t(80) = 2.29$, $p = .012$] with medium effect sizes ($d = 0.51$ to 0.71), and in language skills [composite score, $t(28) = 4.11$, $p < .001$; Phoneme Analysis, $t(29) = 3.75$, $p < .001$; Language Expressive, $t(28) = 3.31$, $p = .001$; Language Receptive, $t(29) = 4.52$, $p < .001$] with large effect sizes ($d = 1.22$ to 1.63). Furthermore, autistic participants showed significantly lower group mean values than the control sample for the evaluation of participation during the test session of intelligence, $t(81) = 2.68$, $p = .004$, executive functions, $t(71) = 2.13$, $p = .018$, and developmental functions, $t(79) = 3.30$, $p < .001$, with medium effect sizes ($d = 0.50$ to 0.73). We found no differences in the subtests Logical–Mathematical Reasoning, $t(81) = 1.44$, $p = .153$, Reading, $t(74) = 1.35$, $p = .182$, Spelling, $t(65) = 1.26$, $p = .212$, and in the motivation and attitude domain [composite score, $t(46) = 0.11$, $p = .458$; Conscientiousness, $t(45) = 0.06$, $p = .477$; Achievement Motivation, $t(47) = -0.07$, $p = .528$], indicating similar performance in autistic and non-autistic participants.

After controlling for multiple testing, significant group differences remained for the composite score of psychomotor skills ($p_H < .001$) and subtests Gross Motor Skills ($p_H < .001$) and Fine Motor Skills ($p_H = .046$). Moreover, the composite score of language skills remained significant ($p_H = .008$) as well as Phoneme Analysis ($p_H = .019$) and Language Receptive ($p_H = .003$) tasks. Finally, the evaluation of participation during testing of the developmental functions remained significant ($p_H = .035$; see Table 4).[3]

3.3. Post Hoc Analyses

To assess for age-related differences between children and adolescents, we further performed post hoc analyses separately for children aged 5–10 years ($n = 17$) and adolescents aged 11–20 years ($n = 26$). After Hommel's (1988) correction, autistic children scored significantly lower than non-autistic children in the composite scores of the cognitive functions, the intelligence group factors, Auditory Short-Term Memory, Visuospatial Short-Term Memory, and Verbal Reasoning (including the corresponding subtests) as well as in psychomotor skills, social–emotional skills, and basic skills of the developmental functions (see Tables S3 and S4 in the Supplemental Material for results). We found no significant group differences between autistic and non-autistic adolescents for the cognitive and developmental functions of the IDS-2 after controlling for multiple testing (see Tables S5 and S6 in the Supplemental Material).

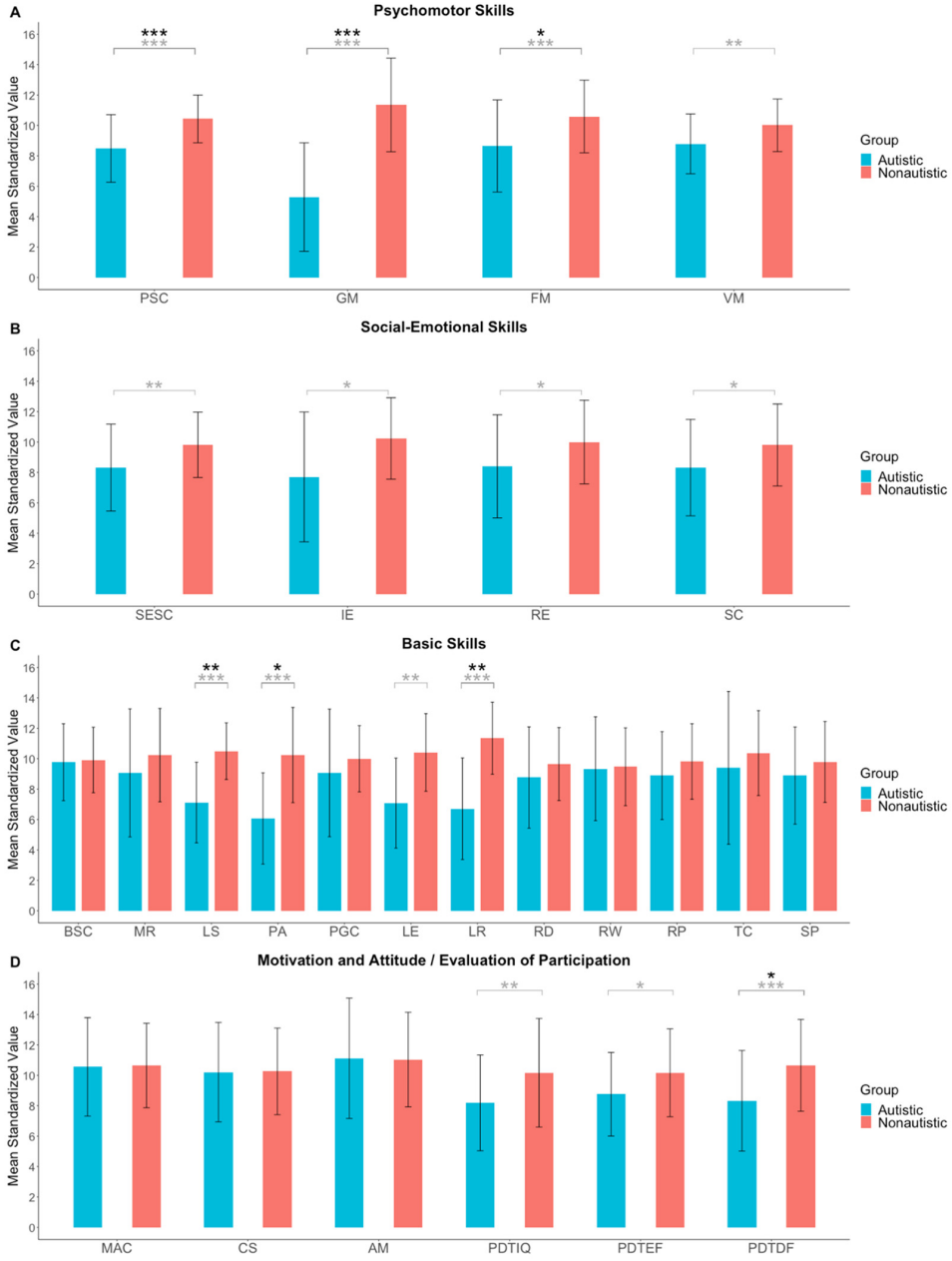

Figure 2. Means and standard deviations are reported for (**A**) psychomotor skills composite score and subtests, (**B**) social–emotional skills composite score and subtests, (**C**) basic skills composite score and subtests, and (**D**) motivation and attitude composite score and subtests as well as for the evaluation of participation during testing of the Intelligence and Development Scales–2 for autistic and non-autistic children and adolescents. Asterisks in grey indicate p values not adjusted with Hommel's (1988) correction. Asterisks in black indicate p values adjusted according to Hommel (1988). PSC = Psychomotor skills composite score; GM = Gross Motor Skills; FM = Fine Motor Skills; VM = Visuomotor Skills; SESC = Social–emotional skills composite score; IE = Identifying Emotions; RE = Regulating Emotions; SC = Socially Competent Behavior; BSC = Basic skills composite score; MR = Logical–Mathematical Reasoning; LS = Language Skills; PA = Phoneme Analysis; PGC = Phoneme–Grapheme Correspondence; LE = Language Expressive; LR = Language Receptive; RD = Reading; RW = Reading Words; RP = Reading Pseudo Words; TC = Text Comprehension; SP = Spelling; MAC = Motivation and attitude composite score; CS = Conscientiousness; AM = Achievement Motivation; PDTIQ = Participation during testing, intelligence; PDTEF = Participation during testing, executive functions; PDTDF = Participation during testing, developmental functions. * $p < .05$. ** $p < .01$. *** $p < .001$.

4. Discussion

In the present study, we compared autistic children and adolescents to a matched control sample on six cognitive and developmental functions assessed with the IDS-2. Our results provide evidence for differential validity for the IDS-2 test scores in psychomotor skills, language skills, and in the evaluation of participation during testing of the developmental functions, with autistic children and adolescents scoring lower than non-autistic participants in these domains. No group differences were detected in the other domains after controlling for multiple testing. Overall, our findings provide an overview of important cognitive and developmental functions in autistic children and adolescents using a single comprehensive and standardized test battery.

In line with our hypotheses, we found similar performance in autistic and non-autistic participants for the intelligence group factors Visual Processing and Abstract Reasoning, which corresponds to studies reporting relative strengths for autistic individuals in nonverbal domains (e.g., Grondhuis et al. 2018) and in subtests measuring fluid reasoning and visuospatial abilities (Charman et al. 2011; Nader et al. 2016). Specifically, the Shape Design subtest, which is part of the Visual Processing group factor of the IDS-2, requires participants to reproduce presented geometric figures with rectangles and triangles. This task is similar to the Block Design subtest of the Wechsler Intelligence Scales, for which autistic individuals oftentimes show at least comparable performance to non-autistic controls (e.g., Muth et al. 2014).

However, in contrast to our hypotheses and previous research (e.g., Demetriou et al. 2018), no significant group differences emerged for the other cognitive functions scores of the IDS-2 after correcting for multiple testing, even though effect sizes were in the small-to-medium range. This finding suggests that our autistic sample included participants with overall average cognitive abilities. One explanation for this result could be that about half of our autistic participants had been diagnosed with Asperger's syndrome, which is known for impairments in social interaction and restricted interests, but without deficits in cognitive development (10th ed.; World Health Organization 2016). Moreover, when assessing age-related differences in a set of post hoc analyses, we found that autistic adolescents scored similarly to non-autistic adolescents in the IDS-2, while autistic children obtained significantly lower scores in several domains of the IDS-2 compared to non-autistic children. In particular, group differences between autistic and non-autistic children remained significant after controlling for multiple testing in the composite scores of the intelligence and executive functions domains as well as in the intelligence group factors Verbal Reasoning and Auditory and Visuospatial Short-Term Memory. These results are in line with previous research reporting weaknesses of autistic children in verbal domains (e.g., Oliveras-Rentas et al. 2012) and in working memory tasks (e.g., Mayes and Calhoun 2003a) as the IDS-2 Auditory and Visuospatial Short-Term Memory group factors also include tasks measuring working memory (i.e., [Mixed] Digit and Letter Span—backwards and Rotated Shape Memory; see Table S1 in the Supplement). In addition, autistic children scored lower on motor and language skills, and importantly, also on social–emotional skills. Interestingly, we did not find any differences between autistic and non-autistic participants when focusing on adolescents only. One reason for this finding could be that autistic adolescents have already received support and intervention in crucial developmental areas, whereas the included autistic children may have been recently diagnosed with autism and thus have had little or no treatment to that point. However, it should be noted that these results are based on small sample sizes. Thus, future studies should use larger age-specific samples to investigate developmental effects across childhood and adolescence and simultaneously control for previous interventions.

Autistic participants had significant impairments in overall psychomotor skills as well as lower scores in gross and fine motor skills in the IDS-2 compared to the non-autistic participants. This finding is in line with results of a previous meta-analysis (Coll et al. 2020) and studies using the M-ABC-2 to assess motor abilities (e.g., Manicolo et al. 2019). Motor skills are particularly important for carrying out everyday tasks (e.g., grasping a glass) and

performing activities of daily living (MacDonald et al. 2013a), as well as for participating in activities at school or in the community (Oliveira et al. 2021). It has been suggested that one reason for these motor differences may be that autistic individuals encounter problems in the translation of sensory inputs into movements (Hannant et al. 2016). Moreover, structural and functional alterations in motor cortex regions of the brain (Mostofsky et al. 2007; Nebel et al. 2014) and in the cerebellum (Fatemi et al. 2012; Mostofsky et al. 2009) have been detected for autistic individuals, which might explain some of the motor impairments. The strong group difference we observed in gross motor skills, representing the largest effect in our study, is in accordance with previous research (Coll et al. 2020) and may be associated with the high prevalence of autistic individuals exhibiting hypotonia (51%) or motor apraxia (34%; Ming et al. 2007). Hence, autistic individuals tend to experience difficulties especially in movements that require activation of muscles in the entire body including balance, arm movements, and coordination. However, as this subtest is administered only to 5- to 10-year-olds in the IDS-2 and correlational research has shown that autistic children's motor skills improve with age (Coll et al. 2020), future longitudinal studies are needed to study possible developmental effects. Although it is not compulsory to report potential difficulties in motor skills as part of the diagnostic criteria of ASD, our findings support the importance of assessing psychomotor abilities during the diagnostic evaluation of children and adolescents at increased likelihood of ASD, as they might be crucial for treatment programs (Bhat et al. 2011; Colombo-Dougovito and Block 2019).

As stated in previous studies, we found that autistic children scored lower in language skills, such as in phoneme analysis (Dynia et al. 2019) and receptive language tasks (Kwok et al. 2015), compared to the non-autistic participants. However, we detected no significant group differences after correcting for multiple testing in expressive language tasks. Although a previous meta-analysis showed equally impaired receptive and expressive language skills in autistic individuals (Kwok et al. 2015), our finding is in line with other studies that also indicated an atypical language pattern of autistic individuals with an advantage in expressive over receptive language skills (e.g., Hudry et al. 2010). One reason for this result might be that we used a direct measurement of language skills in our study. Previous research also found this pattern when using a similar test procedure but did not detect any expressive language advantages when using caregiver reports (Ellis Weismer et al. 2010). Given that having better language production than comprehension skills is contrary to what is generally anticipated in typically developing peers, researchers even suggested that this pattern may be unique to autism (e.g., Volden et al. 2011) and therefore could be used for differential diagnosis (Mitchell et al. 2011) and specific interventions (Hudry et al. 2010). Nevertheless, as the expressive and receptive language tasks are conducted only with 5- to 10-year-olds in the IDS-2 and previous studies have reported a decrease in the expressive–receptive discrepancy in older autistic individuals (Kwok et al. 2015; Volden et al. 2011), it could also be that our result was driven by age effects. Because of the diagnostic and therapeutic potential of this finding, future studies should continue to examine this potential discrepancy between expressive and receptive language in autistic individuals across development.

Additionally, we found no significant group differences in tasks measuring phoneme–grapheme correspondence, which is consistent with our finding that autistic participants also scored similarly to the non-autistic control group in the reading and spelling subtests in our study. This result might be explained by the fact that knowledge of letter–sound correspondence is a prerequisite for the development of literacy skills (Carnine et al. 2010) and therefore needs to be intact for average reading and spelling skills. The finding that our autistic participants showed no differences in the basic skills logical–mathematical reasoning, reading, and spelling compared to non-autistic peers is in line with other studies (e.g., Brown et al. 2013; Chiang and Lin 2007). One reason may refer to the fact that most of the autistic participants in our study attended inclusive educational settings. The enrollment in integrative settings can have a positive impact on autistic individuals' academic skills as individualized education plans in mainstream programs focus more

on academic enhancement than in specialized settings which place more emphasis on life competencies and developmental domains (Kurth and Mastergeorge 2010).

Contrary to previous research (e.g., Cai et al. 2018; Yeung 2022), we found no significant group differences for social–emotional skills after correcting for multiple testing. One explanation for this result could be that the tasks assessing social–emotional skills in the IDS-2 mainly measure explicit knowledge, such as naming socially competent behavior in hypothetical social situations, rather than actual behavior in real-life situations. Since we did not observe any group differences in the cognitive functions of the IDS-2 either, it might be that autistic participants could compensate for difficulties in social–emotional skills with higher-level analytical strategies (Harms et al. 2010; Leung et al. 2022). This would be in line with studies reporting that intelligence is positively associated with social–emotional skills (Jones et al. 2011), especially in autistic individuals (Dyck et al. 2006; Salomone et al. 2019; Trevisan and Birmingham 2016). We found further evidence for this assumption in supplementary analyses where we matched the non-autistic control sample by age, sex, and Full-Scale IQ and obtained lower effect sizes for the social–emotional skills composite score as well as for the subtests Identifying Emotions and Regulating Emotions compared to the effect sizes obtained by matching the samples by age, sex, and maternal education (see Table S2 in the Supplemental Material). In addition, time limits in testing procedures might explain part of the nonsignificant group differences in social–emotional skills. Nagy et al. (2021) found impairments only when time limits for responding were applied, and the present tasks assessing social–emotional skills did not have any time restrictions. However, it is important to note that although meta-analyses and reviews show significant deficits in social–emotional abilities of autistic individuals (e.g., Cai et al. 2018; Yeung 2022), several previous studies were also not able to detect impairments in emotion recognition and regulation (e.g., Jones et al. 2011; Mazefsky et al. 2014; Rosset et al. 2008) or reported difficulties only for certain emotions, for example, for negative emotions (e.g., Shanok et al. 2019). To clarify the interplay between explicit knowledge and social–emotional skills in the IDS-2, future research should use multiple methods to assess social–emotional skills and compare the autistic participants' performance in the IDS-2 with the behavior they demonstrate in real-life social interactions using observational measures. Even though the group differences in the social–emotional skills of the IDS-2 were no longer significant after correcting for multiple testing, it is crucial to mention that effect sizes were within a medium range and comparable to those in a previous meta-analysis (Yeung 2022) which at least tends to indicate differential validity of test scores from the social–emotional skills domain of the IDS-2.

A strength of our study is that we assessed the cognitive and developmental functions using a standardized test procedure with good psychometric properties. Moreover, we used a single test battery based on one standardization sample for the assessment of a broad range of cognitive and developmental domains. In addition, our sample covered a wide age range and was representative of the autistic population, in that the male:female ratio was approximately 4:1 (Maenner et al. 2020), different subtypes were included, and children and adolescents exhibited known comorbid conditions (Leyfer et al. 2006; Salazar et al. 2015). We also consider it a strength that we included participants with intellectual functioning below 70, which represents an understudied subpopulation in autism research (Russell et al. 2019). In addition, by selecting the control sample through a matching procedure, we could control for possible confounding influences of age, sex, and SES.

The present study also has limitations that need to be considered and addressed in future research. First, we relied on diagnostic evaluations carried out by clinical services and experienced psychiatrists and psychotherapists and hence could not consider the standardization and comparability of the diagnoses. Second, we had no information regarding symptom severity or previous treatment programs and could therefore not control for these factors. Third, analyses were conducted at the group level, which limits generalizability to individuals. Finally, although the sample size was larger than in previous

studies, an even larger sample of children and adolescents would further increase the power to detect small effects in future studies.

5. Conclusions

In sum, our findings suggest that in particular, motor and language skills as well as achievement motivation rated by the test administrator were impaired in autistic children and adolescents in the IDS-2 compared to non-autistic participants, which provides evidence for differential validity for these domains of the IDS-2. The largest difference was found in gross motor skills. We therefore advise that therapists working with autistic children should gain knowledge in the area of motor and language therapeutic intervention. Speech–language pathologists as well as psychomotor therapists should obtain autism-specific knowledge, so that autistic children with limited motor and language skills receive appropriate therapeutic support regardless of the background of the therapist. Arguably, with optimal training, autistic participants may also perform tasks in the psychomotor and language domains with greater engagement, which, in turn, could have a positive impact on the long-term development of their motor and language abilities. In conclusion, our results highlight important domains beyond the core symptoms of ASD that need to be considered in future research, educational contexts, and clinical assessment and that seem particularly critical for interventions.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jintelligence10040112/s1, Table S1: Description of the Composites, Group Factors, and Subtests of the Intelligence and Development Scales–2, Table S2: Means, Standard Deviations, and *t* tests of the Developmental Functions From the Intelligence and Development Scales–2 for Autistic and Non-Autistic Children and Adolescents Matched by Age, Sex, and Intelligence, Table S3: Means, Standard Deviations, and *t* tests of the Cognitive Functions From the Intelligence and Development Scales–2 for Autistic and Non-Autistic Children (Aged 5–10 Years), Table S4: Means, Standard Deviations, and *t* tests of the Developmental Functions From the Intelligence and Development Scales–2 for Autistic and Non-Autistic Children (Aged 5–10 Years), Table S5: Means, Standard Deviations, and *t* tests of the Cognitive Functions From the Intelligence and Development Scales–2 for Autistic and Non-Autistic Adolescents (Aged 11–20 Years), Table S6: Means, Standard Deviations, and *t* tests of the Developmental Functions From the Intelligence and Development Scales–2 for Autistic and Non-Autistic Adolescents (Aged 11–20 Years).

Author Contributions: Conceptualization, S.D.O., W.M., S.G. and A.G.; methodology, S.D.O., W.M. and S.G.; software, S.D.O.; validation, S.D.O., W.M. and S.G.; formal analysis, S.D.O.; investigation, S.D.O.; resources, A.G.; data curation, S.D.O.; writing—original draft preparation, S.D.O.; writing—review and editing, S.D.O., W.M., S.G. and A.G.; visualization, S.D.O.; supervision, A.G.; project administration, A.G.; funding acquisition, A.G. 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 Committee Northwest and Central Switzerland (protocol code: PB_2016-01836 and date of approval: 2 May 2019).

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. The data are not publicly available due to privacy issues and property rights.

Acknowledgments: We would like to thank all the children and adolescents for their participation and all the test administrators, especially Miriam Weibel, for their support during data collection. We thank Anita Todd for copyediting the manuscript.

Conflicts of Interest: The authors declare no conflict of interest, with one exception: A.G. is recipient of royalties for the Intelligence and Development Scales–2 (IDS-2).

Notes

[1] According to current models of intelligence (Schneider and McGrew 2018) and executive functions (Miyake et al. 2000), working memory can be understood as a component of intelligence or executive functions. Because working memory is included in the intelligence domain in the IDS-2, we subsumed working memory under the realm of intelligence.

[2] Although the sample size met the robustness criteria for using independent-samples *t* tests (Eid et al. 2017), we also examined the variables regarding normal distribution and variance homogeneity. Analyses using the Shapiro–Wilk test showed that 12 of the 55 dependent variables may not fulfill the normality assumption. Therefore, we additionally calculated Mann–Whitney *U* tests for these variables. The results remained largely the same with two exceptions: First, the mean difference in the subtest Identifying Emotions was no longer significant before controlling for multiple testing. Second, the mean difference in the composite score of language skills was no longer significant after controlling for multiple testing. Furthermore, we found that the Levene's test was significant for fewer than 10 of the dependent variables, indicating unequal variances. Thus, Welch's *t* tests were additionally performed. The results were identical to those obtained from the independent-samples *t* tests.

[3] To control for effects of intelligence, we repeated the independent-samples *t* tests for the developmental functions with a non-autistic control sample matched by age, sex, and intelligence (Full-Scale IQ). The pattern of results remained largely the same, showing lower group mean values for the autistic participants than for the control sample in the domains psychomotor skills, social–emotional skills, language skills, and participation during testing (see Table S2 in the Supplemental Material for full results). These differences hold when correcting for multiple testing in the domain of psychomotor skills. These post hoc analyses underscore the robustness of our findings.

References

Akshoomoff, Natacha. 2006. Use of the Mullen Scales of Early Learning for the Assessment of Young Children with Autism Spectrum Disorders. *Child Neuropsychology* 12: 269–77. [CrossRef]

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Washington, DC: American Psychiatric Association.

Bhat, Anjana N., Rebecca J. Landa, and James C. (Cole) Galloway. 2011. Current Perspectives on Motor Functioning in Infants, Children, and Adults with Autism Spectrum Disorders. *Physical Therapy* 91: 1116–29. [CrossRef] [PubMed]

Boucher, Jill. 2012. Research Review: Structural Language in Autistic Spectrum Disorder—Characteristics and Causes. *Journal of Child Psychology and Psychiatry* 53: 219–33. [CrossRef] [PubMed]

Bowen, Sonya E. 2014. *Autism Spectrum Disorders (ASD): State of the States of Services and Supports for People with ASD*. Washington, DC: L & M Policy Research.

Brignell, Amanda, Angela T. Morgan, Susan Woolfenden, Felicity Klopper, Tamara May, Vanessa Sarkozy, and Katrina Williams. 2018. A Systematic Review and Meta-Analysis of the Prognosis of Language Outcomes for Individuals with Autism Spectrum Disorder. *Autism & Developmental Language Impairments* 3: 1–19. [CrossRef]

Brown, Heather M., Janis Oram-Cardy, and Andrew Johnson. 2013. A Meta-Analysis of the Reading Comprehension Skills of Individuals on the Autism Spectrum. *Journal of Autism and Developmental Disorders* 43: 932–55. [CrossRef]

Cai, Ru Ying, Amanda L. Richdale, Mirko Uljarević, Cheryl Dissanayake, and Andrea C. Samson. 2018. Emotion Regulation in Autism Spectrum Disorder: Where We Are and Where We Need to Go. *Autism Research* 11: 962–78. [CrossRef]

Carnine, Douglas W., Jerry Silbert, Edward J. Kame'enui, and Sara G. Tarver. 2010. *Direct Instruction Reading*, 5th ed. Columbus: Merrill.

Charman, Tony, Andrew Pickles, Emily Simonoff, Susie Chandler, Tom Loucas, and Gillian Baird. 2011. IQ in Children with Autism Spectrum Disorders: Data from the Special Needs and Autism Project (SNAP). *Psychological Medicine* 41: 619–27. [CrossRef]

Chiang, Hsu-Min, and Yueh-Hsien Lin. 2007. Mathematical Ability of Students with Asperger Syndrome and High-Functioning Autism: A Review of Literature. *Autism* 11: 547–56. [CrossRef]

Cibralic, Sara, Jane Kohlhoff, Nancy Wallace, Catherine McMahon, and Valsamma Eapen. 2019. A Systematic Review of Emotion Regulation in Children with Autism Spectrum Disorder. *Research in Autism Spectrum Disorders* 68: 101422. [CrossRef]

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale: Erlbaum.

Coll, Sarah-Maude, Nicholas E. V. Foster, Alexa Meilleur, Simona M. Brambati, and Krista L. Hyde. 2020. Sensorimotor Skills in Autism Spectrum Disorder: A Meta-Analysis. *Research in Autism Spectrum Disorders* 76: 101570. [CrossRef]

Colombo-Dougovito, Andrew M., and Martin E. Block. 2019. Fundamental Motor Skill Interventions for Children and Adolescents on the Autism Spectrum: A Literature Review. *Review Journal of Autism and Developmental Disorders* 6: 159–71. [CrossRef]

Coolican, Jamesie, Susan E. Bryson, and Lonnie Zwaigenbaum. 2008. Brief Report: Data on the Stanford–Binet Intelligence Scales (5th Ed.) in Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders* 38: 190–97. [CrossRef] [PubMed]

de Vries, Marieke, and Hilde M. Geurts. 2012. Cognitive Flexibility in ASD; Task Switching with Emotional Faces. *Journal of Autism and Developmental Disorders* 42: 2558–68. [CrossRef]

Demetriou, Eleni A., Amit Lampit, Daniel S. Quintana, Sharon L. Naismith, Yun J.C. Song, Julia E. Pye, Ian Hickie, and Adam J. Guastella. 2018. Autism Spectrum Disorders: A Meta-Analysis of Executive Function. *Molecular Psychiatry* 23: 1198–204. [CrossRef] [PubMed]

Dyck, Murray J., Jan P. Piek, David Hay, Leigh Smith, and Joachim Hallmayer. 2006. Are Abilities Abnormally Interdependent in Children With Autism? *Journal of Clinical Child & Adolescent Psychology* 35: 20–33. [CrossRef]

Dynia, Jaclyn M., Allison Bean, Laura M. Justice, and Joan N. Kaderavek. 2019. Phonological Awareness Emergence in Preschool Children with Autism Spectrum Disorder. *Autism & Developmental Language Impairments* 4: 1–15. [CrossRef]

Eid, Michael, Mario Gollwitzer, and Manfred Schmitt. 2017. *Statistik und Forschungsmethoden [Statistics and Research Methods]*, 5th ed. Weinheim: Beltz.

Ellis Weismer, Susan, Catherine Lord, and Amy Esler. 2010. Early Language Patterns of Toddlers on the Autism Spectrum Compared to Toddlers with Developmental Delay. *Journal of Autism and Developmental Disorders* 40: 1259–73. [CrossRef]

Fatemi, S. Hossein, Kimberly A. Aldinger, Paul Ashwood, Margaret L. Bauman, Charles D. Blaha, Gene J. Blatt, Abha Chauhan, Ved Chauhan, Stephen R. Dager, Price E. Dickson, and et al. 2012. Consensus Paper: Pathological Role of the Cerebellum in Autism. *The Cerebellum* 11: 777–807. [CrossRef]

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods* 39: 175–91. [CrossRef]

Finnegan, Elizabeth, and Amy L. Accardo. 2018. Written Expression in Individuals with Autism Spectrum Disorder: A Meta-Analysis. *Journal of Autism and Developmental Disorders* 48: 868–82. [CrossRef]

Georgiou, Alexandra, Spyridon-Georgios Soulis, and Danai Rapti. 2018. Motivation in Mathematics of High Functioning Students With Autism Spectrum Disorder (ASD). *Journal of Psychology Research* 8: 96–106. [CrossRef]

Grieder, Silvia, and Alexander Grob. 2020. Exploratory Factor Analyses of the Intelligence and Development Scales–2: Implications for Theory and Practice. *Assessment* 27: 1853–69. [CrossRef]

Grob, Alexander, and Priska Hagmann-von Arx. 2018a. *Intelligence and Development Scales–2 (IDS-2). Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche [Intelligence and Development Scales for Children and Adolescents]*. Bern: Hogrefe.

Grob, Alexander, and Priska Hagmann-von Arx. 2018b. *Intelligence and Development Scales–2 (IDS-2). Intelligenz- und Entwicklungsskalen für Kinder und Jugendliche. Manual zu Theorie, Interpretation und Gütekriterien [Intelligence and Development Scales for Children and Adolescents. Manual on Theory, Interpretation and Psychometric Criteria]*. Bern: Hogrefe.

Grob, Alexander, Priska Hagmann-von Arx, Aleksandra Jaworowska, Anna Matczak, and Diana Fecenec. 2019. *Intelligence and Development Scales-2. Inteligencji i Rozwoju Dla Dzieci i Młodzieży [Intelligence and Development Scales for Children and Adolescents]*. Warsaw: Hogrefe.

Grob, Alexander, Priska Hagmann-von Arx, Anna Barnett, Nichola Stuart, and Serena Vanzan. 2021. *Intelligence and Development Scales-2 (IDS-2). Intelligence and Development Scales for Children and Adolescents*. Oxford: Hogrefe.

Grob, Alexander, Priska Hagmann-von Arx, Rosa Ferri, Monica Rea, and Maria Casagrande. 2022. *Intelligence and Development Scales-2. Scale Di Intelligenza e Sviluppo per Bambini e Adolescenti [Intelligence and Development Scales for Children and Adolescents]*. Florence: Hogrefe.

Grob, Alexander, Priska Hagmann-von Arx, Selma A. J. Ruiter, Marieke E. Timmerman, and Linda Visser. 2018. *Intelligence and Development Scales–2 (IDS-2). Intelligentie- En Ontwikkelingsschalen Voor Kinderen En Jongeren. [Intelligence and Development Scales for Children and Adolescents]*. Amsterdam: Hogrefe.

Grondhuis, Sabrina N., Luc Lecavalier, L. Eugene Arnold, Benjamin L. Handen, Lawrence Scahill, Christopher J. McDougle, and Michael G. Aman. 2018. Differences in Verbal and Nonverbal IQ Test Scores in Children with Autism Spectrum Disorder. *Research in Autism Spectrum Disorders* 49: 47–55. [CrossRef]

Hannant, Penelope, Teresa Tavassoli, and Sarah Cassidy. 2016. The Role of Sensorimotor Difficulties in Autism Spectrum Conditions. *Frontiers in Neurology* 7: 124. [CrossRef] [PubMed]

Happé, Francesca, and Uta Frith. 2020. Annual Research Review: Looking Back to Look Forward—Changes in the Concept of Autism and Implications for Future Research. *Journal of Child Psychology and Psychiatry* 61: 218–32. [CrossRef] [PubMed]

Harms, Madeline B., Alex Martin, and Gregory L. Wallace. 2010. Facial Emotion Recognition in Autism Spectrum Disorders: A Review of Behavioral and Neuroimaging Studies. *Neuropsychology Review* 20: 290–322. [CrossRef]

Hill, Elisabeth L. 2004. Executive Dysfunction in Autism. *Trends in Cognitive Sciences* 8: 26–32. [CrossRef]

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2011. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* 42: 1–62. [CrossRef]

Hommel, Gerhard. 1988. A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Test. *Biometrika* 75: 383–86. [CrossRef]

Howlin, Patricia, and Philippa Moss. 2012. Adults with Autism Spectrum Disorders. *The Canadian Journal of Psychiatry* 57: 275–83. [CrossRef]

Hudry, Kristelle, Kathy Leadbitter, Kathryn Temple, Vicky Slonims, Helen McConachie, Catherine Aldred, Patricia Howlin, Tony Charman, and the PACT Consortium. 2010. Preschoolers with Autism Show Greater Impairment in Receptive Compared with Expressive Language Abilities. *International Journal of Language & Communication Disorders* 45: 681–90. [CrossRef]

Idring, Selma, Michael Lundberg, Harald Sturm, Christina Dalman, Clara Gumpert, Dheeraj Rai, Brian K. Lee, and Cecilia Magnusson. 2015. Changes in Prevalence of Autism Spectrum Disorders in 2001–2011: Findings from the Stockholm Youth Cohort. *Journal of Autism and Developmental Disorders* 45: 1766–73. [CrossRef]

Jahromi, Laudan B., Shantel E. Meek, and Sharman Ober-Reynolds. 2012. Emotion Regulation in the Context of Frustration in Children with High Functioning Autism and Their Typical Peers. *Journal of Child Psychology and Psychiatry* 53: 1250–58. [CrossRef] [PubMed]

Jones, Catherine R. G., Andrew Pickles, Milena Falcaro, Anita J. S. Marsden, Francesca Happé, Sophie K. Scott, Disa Sauter, Jenifer Tregay, Rebecca J. Phillips, Gillian Baird, and et al. 2011. A Multimodal Approach to Emotion Recognition Ability in Autism Spectrum Disorders. *Journal of Child Psychology and Psychiatry* 52: 275–85. [CrossRef] [PubMed]

Keen, Deb. 2009. Engagement of Children with Autism in Learning. *Australasian Journal of Special Education* 33: 130–40. [CrossRef]

Keen, Deb, Amanda Webster, and Greta Ridley. 2016. How Well Are Children with Autism Spectrum Disorder Doing Academically at School? An Overview of the Literature. *Autism* 20: 276–94. [CrossRef] [PubMed]

Kjellmer, Liselotte, Elisabeth Fernell, Christopher Gillberg, and Fritjof Norrelgen. 2018. Speech and Language Profiles in 4- to 6-Year-Old Children with Early Diagnosis of Autism Spectrum Disorder without Intellectual Disability. *Neuropsychiatric Disease and Treatment* 14: 2415–27. [CrossRef]

Konstantareas, M. Mary, and Kelly Stewart. 2006. Affect Regulation and Temperament in Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders* 36: 143–54. [CrossRef]

Kurth, Jennifer A., and Ann M. Mastergeorge. 2010. Academic and Cognitive Profiles of Students with Autism: Implications for Classroom Practice and Placement. *International Journal of Special Education* 25: 8–14.

Kwok, Elaine Y. L., Heather M. Brown, Rachael E. Smyth, and Janis Oram Cardy. 2015. Meta-Analysis of Receptive and Expressive Language Skills in Autism Spectrum Disorder. *Research in Autism Spectrum Disorders* 9: 202–22. [CrossRef]

Leung, Florence Yik Nam, Jacqueline Sin, Caitlin Dawson, Jia Hoong Ong, Chen Zhao, Anamarija Veić, and Fang Liu. 2022. Emotion Recognition across Visual and Auditory Modalities in Autism Spectrum Disorder: A Systematic Review and Meta-Analysis. *Developmental Review* 63: 101000. [CrossRef]

Leyfer, Ovsanna T., Susan E. Folstein, Susan Bacalman, Naomi O. Davis, Elena Dinh, Jubel Morgan, Helen Tager-Flusberg, and Janet E. Lainhart. 2006. Comorbid Psychiatric Disorders in Children with Autism: Interview Development and Rates of Disorders. *Journal of Autism and Developmental Disorders* 36: 849–61. [CrossRef] [PubMed]

Libertus, Klaus, Kelly A. Sheperd, Samuel W. Ross, and Rebecca J. Landa. 2014. Limited Fine Motor and Grasping Skills in 6-Month-Old Infants at High Risk for Autism. *Child Development* 85: 2218–31. [CrossRef] [PubMed]

Lienert, Gustav A., and Ulrich Raatz. 1998. *Testaufbau und Testanalyse [Test Design and Test Analysis]*. Weinheim: Beltz.

Liu, Ting, and Casey M. Breslin. 2013. Fine and Gross Motor Performance of the MABC-2 by Children with Autism Spectrum Disorder and Typically Developing Children. *Research in Autism Spectrum Disorders* 7: 1244–49. [CrossRef]

Lodi-Smith, Jennifer, Jonathan D. Rodgers, Sara A. Cunningham, Christopher Lopata, and Marcus L. Thomeer. 2019. Meta-Analysis of Big Five Personality Traits in Autism Spectrum Disorder. *Autism* 23: 556–65. [CrossRef]

Luyster, Rhiannon, Kristina Lopez, and Catherine Lord. 2007. Characterizing Communicative Development in Children Referred for Autism Spectrum Disorders Using the MacArthur-Bates Communicative Development Inventory (CDI). *Journal of Child Language* 34: 623–54. [CrossRef]

MacDonald, Megan, Catherine Lord, and Dale A. Ulrich. 2013a. The Relationship of Motor Skills and Adaptive Behavior Skills in Young Children with Autism Spectrum Disorders. *Research in Autism Spectrum Disorders* 7: 1383–90. [CrossRef]

MacDonald, Megan, Catherine Lord, and Dale A. Ulrich. 2013b. The Relationship of Motor Skills and Social Communicative Skills in School-Aged Children with Autism Spectrum Disorder. *Adapted Physical Activity Quarterly* 30: 271–82. [CrossRef]

Maenner, Matthew J., Kelly A. Shaw, Jon Baio, Anita Washington, Mary Patrick, Monica DiRienzo, Deborah L. Christensen, Lisa D. Wiggins, Sydney Pettygrove, Jennifer G. Andrews, and et al. 2020. Prevalence of Autism Spectrum Disorder among Children Aged 8 Years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016. *MMWR Surveillance Summaries* 69: 1–12. [CrossRef]

Mandelbaum, David E., Michael Stevens, Eric Rosenberg, Max Wiznitzer, Mitchell Steinschneider, Saul R. Korey, Pauline Filipek, Isabelle Rapin, and Saul R. Korey. 2006. Sensorimotor Performance in School-Age Children with Autism, Developmental Language Disorder, or Low IQ. *Developmental Medicine & Child Neurology* 48: 33–39. [CrossRef]

Manicolo, Olivia, Mark Brotzmann, Priska Hagmann-von Arx, Alexander Grob, and Peter Weber. 2019. Gait in Children with Infantile/Atypical Autism: Age-Dependent Decrease in Gait Variability and Associations with Motor Skills. *European Journal of Paediatric Neurology* 23: 117–25. [CrossRef] [PubMed]

Mayes, Susan Dickerson, and Susan L. Calhoun. 2003a. Analysis of WISC-III, Stanford-Binet:IV, and Academic Achievement Test Scores in Children with Autism. *Journal of Autism and Developmental Disorders* 33: 329–41. [CrossRef] [PubMed]

Mayes, Susan Dickerson, and Susan L. Calhoun. 2003b. Ability Profiles in Children with Autism: Influence of Age and IQ. *Autism* 7: 65–80. [CrossRef] [PubMed]

Mazefsky, Carla A., Xenia Borue, Taylor N. Day, and Nancy J. Minshew. 2014. Emotion Regulation Patterns in Adolescents with High-Functioning Autism Spectrum Disorder: Comparison to Typically Developing Adolescents and Association with Psychiatric Symptoms. *Autism Research* 7: 344–54. [CrossRef]

Meyer, Christine Sandra, Priska Hagmann-von Arx, and Alexander Grob. 2009. Die Intelligence and Development Scale Sozial-Emotionale Kompetenz (IDS-SEK): Psychometrische Eigenschaften eines Tests zur Erfassung sozial-emotionaler Fähigkeiten [The Intelligence and Development Scale Social-Emotional Competence (IDS-SEK): Psychometric properties of a test to assess social-emotional skills]. *Diagnostica* 55: 234–44. [CrossRef]

Ming, Xue, Michael Brimacombe, and George C. Wagner. 2007. Prevalence of Motor Impairment in Autism Spectrum Disorders. *Brain and Development* 29: 565–70. [CrossRef]

Mitchell, Shelley, Janis Oram Cardy, and Lonnie Zwaigenbaum. 2011. Differentiating Autism Spectrum Disorder from Other Developmental Delays In The First Two Years Of Life. *Developmental Disabilities Research Reviews* 17: 130–40. [CrossRef]

Mitchell, Shelley, Jessica Brian, Lonnie Zwaigenbaum, Wendy Roberts, Peter Szatmari, Isabel Smith, and Susan Bryson. 2006. Early Language and Communication Development of Infants Later Diagnosed with Autism Spectrum Disorder. *Journal of Developmental & Behavioral Pediatrics* 27: S69–S78.

Miyake, Akira, and Naomi P. Friedman. 2012. The Nature and Organization of Individual Differences in Executive Functions: Four General Conclusions. *Current Directions in Psychological Science* 21: 8–14. [CrossRef]

Miyake, Akira, Naomi P. Friedman, Michael J. Emerson, Alexander H. Witzki, Amy Howerter, and Tor D. Wager. 2000. The Unity and Diversity of Executive Functions and Their Contributions to Complex 'Frontal Lobe' Tasks: A Latent Variable Analysis. *Cognitive Psychology* 41: 49–100. [CrossRef]

Mostofsky, Stewart H., Melanie P. Burgess, and Jennifer C. Gidley Larson. 2007. Increased Motor Cortex White Matter Volume Predicts Motor Impairment in Autism. *Brain* 130: 2117–22. [CrossRef] [PubMed]

Mostofsky, Stewart H., Stephanie K. Powell, Daniel J. Simmonds, Melissa C. Goldberg, Brian Caffo, and James J. Pekar. 2009. Decreased Connectivity and Cerebellar Activity in Autism during Motor Task Performance. *Brain* 132: 2413–25. [CrossRef] [PubMed]

Muth, Anne, Johannes Hönekopp, and Christine M. Falter. 2014. Visuo-Spatial Performance in Autism: A Meta-Analysis. *Journal of Autism and Developmental Disorders* 44: 3245–63. [CrossRef] [PubMed]

Nader, Anne-Marie, Valérie Courchesne, Michelle Dawson, and Isabelle Soulières. 2016. Does WISC-IV Underestimate the Intelligence of Autistic Children? *Journal of Autism and Developmental Disorders* 46: 1582–89. [CrossRef]

Nagy, Emese, Louise Prentice, and Tess Wakeling. 2021. Atypical Facial Emotion Recognition in Children with Autism Spectrum Disorders: Exploratory Analysis on the Role of Task Demands. *Perception* 50: 819–33. [CrossRef]

Nebel, Mary Beth, Suresh E. Joel, John Muschelli, Anita D. Barber, Brian S. Caffo, James J. Pekar, and Stewart H. Mostofsky. 2014. Disruption of Functional Organization within the Primary Motor Cortex in Children with Autism. *Human Brain Mapping* 35: 567–80. [CrossRef]

Oliveira, Katherine Simone Caires, Déborah Ebert Fontes, Egmar Longo, Hércules Ribeiro Leite, and Ana Cristina Resende Camargos. 2021. Motor Skills Are Associated with Participation of Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders* 21: 1–10. [CrossRef]

Oliveras-Rentas, Rafael E., Lauren Kenworthy, Richard B. Roberson, Alex Martin, and Gregory L. Wallace. 2012. WISC-IV Profile in High-Functioning Autism Spectrum Disorders: Impaired Processing Speed Is Associated with Increased Autism Communication Symptoms and Decreased Adaptive Communication Abilities. *Journal of Autism and Developmental Disorders* 42: 655–64. [CrossRef]

Pennington, Bruce F., and Sally Ozonoff. 1996. Executive Functions and Developmental Psychopathology. *Journal of Child Psychology and Psychiatry* 37: 51–87. [CrossRef]

Petermann, Franz. 2008. *Movement Assessment Battery for Children*, 2nd ed. Frankfurt: Pearson Assessment.

Provost, Beth, Brian R. Lopez, and Sandra Heimerl. 2007. A Comparison of Motor Delays in Young Children: Autism Spectrum Disorder, Developmental Delay, and Developmental Concerns. *Journal of Autism and Developmental Disorders* 37: 321–28. [CrossRef]

R Core Team. 2021. *R: A Language and Environment for Statistical Computing (Version 4.0.3) [Computer Software]*. Vienna: R Foundation for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 13 February 2021).

Rosset, Delphine B., Cécilie Rondan, David Da Fonseca, Andreia Santos, Brigitte Assouline, and Christine Deruelle. 2008. Typical Emotion Processing for Cartoon but Not for Real Faces in Children with Autistic Spectrum Disorders. *Journal of Autism and Developmental Disorders* 38: 919–25. [CrossRef] [PubMed]

Russell, Ginny, William Mandy, Daisy Elliott, Rhianna White, Tom Pittwood, and Tamsin Ford. 2019. Selection Bias on Intellectual Ability in Autism Research: A Cross-Sectional Review and Meta-Analysis. *Molecular Autism* 10: 9. [CrossRef] [PubMed]

Salazar, Fernando, Gillian Baird, Susie Chandler, Evelin Tseng, Tony O'sullivan, Patricia Howlin, Andrew Pickles, and Emily Simonoff. 2015. Co-Occurring Psychiatric Disorders in Preschool and Elementary School-Aged Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders* 45: 2283–94. [CrossRef] [PubMed]

Salomone, Erica, Daniela Bulgarelli, Evelyne Thommen, Emanuelle Rossini, and Paola Molina. 2019. Role of Age and IQ in Emotion Understanding in Autism Spectrum Disorder: Implications for Educational Interventions. *European Journal of Special Needs Education* 34: 383–92. [CrossRef]

Schmidt-Atzert, Lothar, and Manfred Amelang. 2012. *Psychologische Diagnostik [Psychological Assessment]*, 5th ed. Berlin: Springer.

Schneider, W. Joel, and Kevin S. McGrew. 2018. The Cattell–Horn–Carroll Theory of Cognitive Abilities. In *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, 4th ed. Edited by Dawn P. Flanagan and Erin M. McDonough. New York: Guilford Press, pp. 73–163.

Shanok, Nathaniel A., Nancy Aaron Jones, and Nikola N. Lucas. 2019. The Nature of Facial Emotion Recognition Impairments in Children on the Autism Spectrum. *Child Psychiatry & Human Development* 50: 661–67. [CrossRef]

Siaperas, Panagiotis, Howard A. Ring, Catherine J. McAllister, Sheila Henderson, Anna Barnett, Peter Watson, and Anthony J. Holland. 2012. Atypical Movement Performance and Sensory Integration in Asperger's Syndrome. *Journal of Autism and Developmental Disorders* 42: 718–25. [CrossRef]

Thomas, Pauline, Walter Zahorodny, Bo Peng, Soyeon Kim, Nisha Jani, William Halperin, and Michael Brimacombe. 2012. The Association of Autism Diagnosis with Socioeconomic Status. *Autism* 16: 201–13. [CrossRef]

Titeca, Daisy, Herbert Roeyers, and Annemie Desoete. 2017. Early Numerical Competencies in 4- and 5-Year-Old Children with Autism Spectrum. *Focus on Autism and Other Developmental Disabilities* 32: 279–92. [CrossRef]

Trevisan, Dominic A., and Elina Birmingham. 2016. Are Emotion Recognition Abilities Related to Everyday Social Functioning in ASD? A Meta-Analysis. *Research in Autism Spectrum Disorders* 32: 24–42. [CrossRef]

Troyb, Eva, Alyssa Orinstein, Katherine Tyson, Molly Helt, Inge-Marie Eigsti, Michael Stevens, and Deborah Fein. 2014. Academic Abilities in Children and Adolescents with a History of Autism Spectrum Disorders Who Have Achieved Optimal Outcomes. *Autism* 18: 233–43. [CrossRef]

Van Meter, Karla C., Lasse E. Christiansen, Lora D. Delwiche, Rahman Azari, Tim E. Carpenter, and Irva Hertz-Picciotto. 2010. Geographic Distribution of Autism in California: A Retrospective Birth Cohort Analysis. *Autism Research* 3: 19–29. [CrossRef] [PubMed]

Volden, Joanne, Isabel M. Smith, Peter Szatmari, Susan Bryson, Eric Fombonne, Pat Mirenda, Wendy Roberts, Tracy Vaillancourt, Charlotte Waddell, Lonnie Zwaigenbaum, and et al. 2011. Using the Preschool Language Scale, Fourth Edition to Characterize Language in Preschoolers with Autism Spectrum Disorders. *American Journal of Speech-Language Pathology* 20: 200–8. [CrossRef]

White, Susan Williams, Lawrence Scahill, Ami Klin, Kathleen Koenig, and Fred R. Volkmar. 2007. Educational Placements and Service Use Patterns of Individuals with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders* 37: 1403–12. [CrossRef] [PubMed]

World Health Organization. 2016. *International Statistical Classification of Diseases and Related Health Problems (10th Rev.)*. Geneva: World Health Organization. Available online: https://icd.who.int/browse10/2019/en (accessed on 5 November 2022).

World Health Organization. 2018. *International Classification of Diseases for Mortality and Morbidity Statistics (11th Rev.)*. Geneva: World Health Organization. Available online: https://icd.who.int/browse11/l-m/en (accessed on 6 January 2022).

Yerys, Benjamin E., Gregory L. Wallace, Jennifer L. Sokoloff, Devon A. Shook, Joette D. James, and Lauren Kenworthy. 2009. Attention Deficit/Hyperactivity Disorder Symptoms Moderate Cognition and Behavior in Children with Autism Spectrum Disorders. *Autism Research* 2: 322–33. [CrossRef] [PubMed]

Yeung, Michael K. 2022. A Systematic Review and Meta-Analysis of Facial Emotion Recognition in Autism Spectrum Disorder: The Specificity of Deficits and the Role of Task Characteristics. *Neuroscience & Biobehavioral Reviews* 133: 104518. [CrossRef]

Zajic, Matthew C., Emily J. Solari, Ryan P. Grimm, Nancy S. McIntyre, and Peter C. Mundy. 2020. Relationships between Reading Profiles and Narrative Writing Abilities in School-Age Children with Autism Spectrum Disorder. *Reading and Writing* 33: 1531–56. [CrossRef]

Journal of
Intelligence

MDPI

Article

Evaluation of the Wechsler Individual Achievement Test-Fourth Edition as a Measurement Instrument

A. Alexander Beaujean [1,*] and Jason R. Parkin [2]

[1] Psychology & Neuroscience Department, Baylor University, Waco, TX 76798-7334, USA
[2] Department of Teaching, Learning and Social Justice, Seattle University, Seattle, WA 98122, USA; parkinj@seattleu.edu
* Correspondence: alex_beaujean@baylor.edu

Abstract: The Wechsler Individual Achievement Test (WIAT-4) is the latest iteration of a popular instrument that psychologists employ to assess academic achievement. The WIAT-4 authors make both pragmatic and measurement claims about the instrument. The pragmatic claims involve being useful for identifying individuals in certain academic achievement-related groups (e.g., specific learning disability). The measurement claims are twofold: (a) the instrument's scores represent psychological attributes, and (b) scores transformed to standard score values have equal-interval properties. The WIAT-4 authors did not provide the evidence necessary to support the pragmatic claims in the technical manual, so we could not evaluate them. Thus, we limited our evaluation to the measurement claims for the composite scores. To do so, we used information in the technical manual along with some additional factor analyses. Support for the first measurement claim varies substantially across scores. Although none of the evidence is particularly strong, scores in mathematics and reading domains tend to have more support than the writing and total achievement scores. Support for the second claim was insufficient for all scores. Consequently, we recommend that psychologists wishing to interpret WIAT-4 composite scores limit those interpretations to just a few in the mathematics and reading domains. Second, psychologists should completely refrain from using any composite score in a way that requires equal-interval values (e.g., quantitative score comparisons). Neither of these recommendations necessarily disqualifies the scores from being useful for pragmatic purposes, but support for these uses will need to come from evidence not currently provided in the WIAT-4 technical manual.

Keywords: validity; Wechsler Individual Achievement Test; test review; measurement; academic achievement

Citation: Beaujean, A. Alexander, and Jason R. Parkin. 2022. Evaluation of the Wechsler Individual Achievement Test-Fourth Edition as a Measurement Instrument. *Journal of Intelligence* 10: 30. https://doi.org/10.3390/jintelligence10020030

Received: 7 January 2022
Accepted: 19 May 2022
Published: 22 May 2022

Users of any psychological instrument have the burden of supporting their use of it (American Educational Research Association et al. 2014; Kline 1998). As such, it is critical that psychologists rigorously evaluate every instrument they employ (Mitchell 1984). It is often years after an instrument is published before peer-reviewed literature is available, so potential users wishing to make an instrument-adoption decision before then must rely on the information produced by the instrument authors.[1] Thus, it is incumbent for instrument authors to provide sufficient information about the instrument for potential users to make an informed decision about whether to adopt the instrument (International Test Commission 2001). In this article, we review the fourth edition of the Wechsler Individual Achievement Test (WIAT-4; NCS Pearson 2020) and evaluate it using information provided in the instrument's technical manual. Before doing so, we first discuss what is involved in evaluating psychological instruments.

1. Evaluating Psychological Instruments

The phrase "evaluating a psychological instrument" is somewhat of a misnomer because it does not involve evaluating an instrument itself as much as it involves evaluating (a) statements (claims) the instrument authors make about its intended uses, and (b)

evidence (arguments) to support the truthfulness of those claims (Campbell et al. 2008; Kane 2013). As such, evaluations of psychological instruments should differ substantially based on the instrument's purposes—something that is often "insufficiently recognized" (Ozer and Reise 1994, p. 363). We can class the purposes for most scientific instruments as measurement or pragmatic (Hand 2016; Lindquist 1936).

Measurement purposes are those that concern representation, specifically depicting an attribute's manifestations and the relations among them (but see Michell 1999). Evaluating measurement claims involves evaluating the instrument's validity (i.e., validation; Borsboom et al. 2004). *Pragmatic purposes* involve making decisions (e.g., provide treatment, make diagnoses), so evaluating pragmatic claims primarily involves evaluating evidence for the scores' utility (e.g., sensitivity, cost-benefit). Pragmatic and measurement purposes are not mutually exclusive, so it is possible to employ an instrument's scores for (a) only pragmatic purposes, (b) only measurement purposes, or (c) both pragmatic and measurement purposes (Newton 2017). Measurement and pragmatic uses are more or less independent of each other, however, so it is possible for an instrument's scores to have strong utility evidence without measuring anything (or vice versa).

1.1. Validity

The concept of validity in the context of psychological measurement goes back to 19th century, but it did not become something of major interest to psychologists until the 20th century (Newton and Shaw 2014). Although validity quickly became an ambiguous concept in psychology (Slaney 2017), since the mid-20th century psychologists have increasingly employed it to mean something external to the instrument and contingent upon on particular interpretations of an instrument's scores (e.g., Guilford 1946; Messick 1989). As such, support for validity claims is viewed as something discoverable through an ongoing process of assessing the correlations between an instrument's scores and other phenomena (Reynolds 1998). This meaning of validity is troublesome (Markus and Borsboom 2013).

Pretend we have an instrument designed to measure people's ability to add integers (i.e., *integer addition*). It may be interesting to know that the instrument's score correlates with scores from other instruments—particularly instruments designed to measure integer addition. Two variables can correlate/not correlate for a variety of reasons,; however, only one of the two involves how well the scores represent integer addition (Borsboom 2005). Moreover, implicit in creating the instrument is some a priori knowledge about the meaning of the integer addition concept as well as the belief that the instrument's score represents that concept (Krause 2005). Thus, correlations themselves cannot be the basis for determining whether the instrument measures integer addition (Guttman 1977, items 30–31). This does *not* entail that empirically acquired information is useless. To the contrary, empirical information is necessary to support certain claims about attributes needed to create a valid instrument (e.g., whether integer addition ability is a quantity; Mari et al. 2015). Likewise, empirical information can aid in selecting items from a pool of potential items that all cohere to the meaning of integer addition (Loevinger 1957) or spur further work in refining the integer addition concept (Krause 2012). Evaluating whether the instrument is valid, however, is a fundamentally a conceptual endeavor.

1.2. Evaluating the Validity of Psychological Instruments

Broadly, scientific instruments have validity to the extent that they measure the attributes they are designed to measure (Joint Committee for Guides in Metrology 2012). This entails that, for an instrument to be valid, (a) the intended-to-measure attribute has to exist as more than just a name (i.e., it has to be potentially measurable); and (b) variation in the attribute impinges on variation in score values the instrument produces (Borsboom et al. 2004). Although relatively straightforward, evaluating validity is not a simple endeavor—especially for instruments measuring psychological attributes. We will discuss a few components to such evaluations.

First, it is necessary to understand the meaning (i.e., rules for employment) of the to-be-measured attribute concept (Michell 2009). Most psychological attribute concepts are *functional*, so their rules for use involve things we do (Bem and De Jong 2013). Only psychological attributes whose meanings involve behavior are open to public observation, so those are the attributes we can ascribe to other people (Bennett and Hacker 2022; Coombs 1948).[2] Being observable does not, however, guarantee measurability. Although we acknowledge there is not currently a consensus about the necessary or sufficient criteria for an attribute to be measurable (Mari et al. 2017), we believe the second and third components we discuss are necessary for measurability.

The second component is understanding how the behaviors that constitute a particular attribute go together. Like all other concepts, psychological attribute concepts are part of language, so psychologists are free to give them whatever meaning they want. As such, the behaviors criterial for a given attribute concept can go together a variety of ways, which are often not obvious. At one extreme are attributes whose behaviors go together because they have functional unity. If behaviors have *functional unity*, then they go together because of the behaviors themselves rather than the meaning of an attribute (Hearnshaw 1941; Peak 1953).[3] In other words, the behaviors would still go together even if the attribute concept did not exist.

At the other extreme are attributes whose behaviors go together by fiat—they go together only because the concept includes them all. For example, psychologists often discuss *job morale* as if it is a single attribute, but the behaviors that constitute it (e.g., initiating activities, not seeking employment elsewhere, few absences) largely only go together because psychologists put them together when defining the job morale concept (Hardy 2009). Thus, it is not uncommon for employees to emit some of the behaviors but not others. To the extent this is true, representing job morale with a single score allows for the possibility of two people to be classified as having equal job morale yet manifest non-overlapping sets of behavior. This makes it difficult to support a claim that job morale is measurable. Two ways to rectify the situation are to restrict use of job morale to a hypernym for classifying job-related behaviors, or to make the meaning and representation of job morale multi-dimensional. Psychologists seldom employ either solution, however, but instead primarily look to study and measure attributes they can represent with a single score (Sijtsma 2006). In such cases, functional unity is a necessary condition for measurement.

Third, it is necessary to know the attribute's different possible manifestations and the relations among the manifestations (e.g., equivalence, order, additivity) because this information determines whether an attribute is a quality, quantity, or something in between (Barrett 2018; Michell 2005). For example, it is self-evident that the integer addition ability has at least two manifestations: can add integers and cannot add integers. People can manifest the ability to add integers different ways, one of which is consistently responding to items about adding integers correctly. Likewise, one way people manifest not having the ability to add integers is consistently responding to integer addition items incorrectly. Since these two manifestations are mutually exclusive (i.e., it would be incoherent to state that the same person can both add integers and not add integers), we can represent the attribute on a so-called nominal scale.[4] Of course, scientists do not rely on intuition for determining the different manifestations of an attribute and their relations. Instead, it is something that requires considerable conceptual and empirical work (Mari et al. 2015; Michell 1990).

Fourth, it is necessary to determine whether the instrument's specifications (e.g., content, procedures) are consistent with what is currently known about the attribute (Krause 1967; Maraun 1998). For example, an instrument would not be valid for measuring the (overly simplistic) integer addition ability if it requires respondents to answer items such as "What is the capital city of Scotland?", but could be valid if it had items such as "2 + 2 = ?". Likewise, instruments producing scores with two values might represent the attribute faithfully (e.g., can/cannot add integers), but instruments producing more than two values (e.g., Normal Curve Equivalents) would not represent the attribute very well. Of course, it is not really the number of possible values that is important, but that all the

known relations among attribute manifestations are faithfully represented in the relations among a score's values.

2. Wechsler Individual Achievement Test–Fourth Edition

The WIAT-4 is multiple things simultaneously. It is (a) a standardized battery of individually administered instruments (i.e., subtests), each of which is comprised of items designed to elicit certain mental attributes and behavior; (b) a set of criteria for coding the elicited behavior; and (c) a set of algorithms for translating the coded behavior into values for different scores (i.e., scoring). As such, it is similar to many other academic achievement instruments currently available (e.g., Bardos 2020; Kaufman et al. 2014). The WIAT-4 is based on the third edition of the instrument (WIAT-3), but it is more than just an updated WIAT-3. The instrument authors not only collected data from a new norming sample, but also substantially added and revised items, subtests, and scores (Breaux 2020, p. 89). In addition, many of WIAT-4 scores are based on a psychological theory, which is notably different from the WIAT-3 wherein all the scores are atheoretical (Breaux 2020, pp. 89–96). As such, it is best to think of the WIAT-4 as a brand-new instrument rather than an update of a previously existing one (Beaujean 2015a; Bush et al. 2018).

2.1. Purpose of Wechsler Individual Achievement Test

The WIAT-4 authors claim the instrument can be used for both measurement and pragmatic purposes. They are explicit in their measurement claims, stating the instrument is "designed to measure the [academic] achievement of examinees ages 4 through 50, and students in prekindergarten (PK) through Grade 12" (Breaux 2020, p. 1; see also p. 28). In addition, the authors state that values of some of the scores "are on an equal-interval scale" (Breaux 2020, p. 64). Evaluating both claims require evaluating (measurement) validity evidence.

The pragmatic purposes involve using WIAT-4 scores for identifying members of various academic achievement-related groups (e.g., gifted, specific learning disability; Breaux 2020, pp. 83–87). Evaluating these claims involves evaluating empirical evidence about the scores' utility. The utility evidence provided in the WIAT-4 technical manual consists of (a) basic descriptive statistics (e.g., means, standard deviation) of the scores for each group; (b) descriptive statistics for between-group score differences (e.g., standardized effect sizes); and (c) *p*-values for null hypotheses regarding mean differences between groups (Breaux 2020, pp. 47–60).[5] While this information is somewhat useful, it is not sufficient for us to evaluate the scores' utility (McFall and Treat 1999). Consequently, in our evaluation we focus exclusively on the evidence supporting the WIAT-4 authors' measurement claims.

2.2. Wechsler Individual Achievement Test Scores

The WIAT-4 produces 32 scores (see Table 1), which we can classify different ways. One classification criterion is whether the score is comprised of other scores. *Simple scores* are those whose values are not dependent on the value of any other scores (i.e., based on a single set of items), while *composite scores* are those whose values are a function of simple scores. All WIAT-4 composite scores are unweighted sums of two or more simple scores (Breaux 2020, pp. 12–13). Most of the WIAT-4 subtests produce simple scores, but there are few exceptions (see notes in Table 1). A second criterion for classing scores is knowledge domain (i.e., content). The WIAT-4 authors designed the subtests' items to elicit abilities in three core academic knowledge domains (i.e., reading, writing, mathematics) as well as in oral language (Breaux 2020, p. 28). All the WIAT-4 scores cover content from a single academic knowledge domain except for two: Total Achievement and Orthographic Processing.

Table 1. Wechsler Individual Achievement Test—Fourth Edition Subtests.

Subtest Scores	Grade Levels	Composite Scores	
		Single Knowledge Domain	**Multiple Knowledge Domains**
		Reading Domain	
Decoding Fluency [a]	3–12+	Reading Fluency (3–12+)	
Oral Reading Fluency	1–12+	Reading Fluency (1–12+)	
Orthographic Fluency [a]	1–12+	Dyslexia Index (4–12+)	Orthographic Processing (1–12+)
		Reading Fluency (1–12+)	
Phonemic Proficiency [a,b]	PK–12+	Basic Reading	
		Dyslexia Index (PK-3)	
		Phonological Processing (1–12+)	
Pseudoword Decoding	1–12+	Basic Reading	
		Decoding	
		Dyslexia Index (4–12+)	
		Phonological Processing (1–12+)	
Reading Comprehension	K-12+	Reading (K-12+)	Total Achievement (PK-12+)
Word Reading	PK-12+	Basic Reading	Total Achievement (PK-12+)
		Decoding	
		Dyslexia Index (PK-12+)	
		Reading (K-12+)	
		Writing Domain	
Alphabet Writing Fluency	PK-4+	Written Expression (K-1)	Total Achievement (PK-1)
		Writing Fluency (1-4)	
Essay Composition	3-12+	Written Expression (4-12+)	Total Achievement (4–12+)
Sentence Composition [c]	1-12+	Written Expression (2-12+)	Total Achievement (2–3)
Sentence Writing Fluency [a]	1-12+	Writing Fluency (1-4)	
Spelling	K-12+	Written Expression (K-12+)	Total Achievement (K-12+)
			Orthographic Processing (1–12+)
		Mathematics Domain	
Math Problem Solving	PK-12+	Mathematics (K-12+)	Total Achievement (PK-12+)
Numerical Operations	K-12+	Mathematics (K-12+)	Total Achievement (K-12+)
Math Fluency–Addition	1–12+	Math Fluency (1–12+)	
Math Fluency–Subtraction	1–12+	Math Fluency (1–12+)	
Math Fluency-Multiplication	3–12+	Math Fluency (3–12+)	
		Oral Language Domain	
Listening Comprehension [d]	PK-12+	Oral Language (PK-12+)	
Oral Expression [e]	PK-12+	Oral Language (PK-12+)	

Note. There is an additional new subtest called *Orthographic Choice*, but it is only available on the Q-Interactive version of the instrument. It combines with the Orthographic Fluency and Spelling subtests to form an Orthographic Processing Extended composite score. [a] Subtest is new to WIAT-IV. [b] Listed as a subtest in the Language Processing domain in technical manual. [c] Listed as a subtest in the technical manual but is comprised of two "component scores:" Sentence Building and Sentence Combining. [d] Listed as a subtest in the technical manual but is comprised of two "component scores:" Receptive Vocabulary and Oral Discourse Comprehension [e]. Listed as a subtest in the technical manual but is comprised of three "component scores:" Expressive Vocabulary, Oral Word Fluency, and Sentence Repetition.

The WIAT-4 authors state that interpreting the WIAT-4 scores should follow a four-step process (Breaux 2020, pp. 77–79).[6]

Step 1. Interpret the Total Achievement score.

Step 2a. Interpret all other composite scores and subtest scores normatively (i.e., compare how a respondent performed in reference to peers of the same age or grade).

Step 2b. Interpret all other composite scores and subtest scores ipsatively (i.e., compare scores within a single respondent).

Step 3. Identify ipsative strengths and weaknesses from composite scores. This involves (a) comparing each single-domain composite score for a respondent to the same respondent's Total Achievement score, and (b) determining if the value difference is statistically different from zero.

Step 4. Make planned ipsative comparisons between different subtest scores or different composite scores. This involves (a) selecting multiple subtest or composite scores to compare, and then (b) determining if their value differences are statistically different from zero.

Implicit in the WIAT-4 interpretive guidance is the claim that each WIAT-4 score represents a distinct, although not necessarily unrelated, attribute. Consequently, it is necessary to evaluate the validity of each score. In this article, we focus on evaluating the evidence for the scores in steps 1 and 2a. We do so for two reasons. First, steps 2b–4 involve ipsative analysis and interpretation. *Ipsative* means "of the self", so steps 2b–4 require comparing scores for a particular respondent to other scores for the same respondent (e.g., compare the Listening Comprehension score to the Reading Comprehension score; Cattell 1944). These interpretations are only warranted if the equal-interval claim is true. Second, although ipsative interpretations require certain measurement properties, they are primarily employed with the WIAT-4 for making pragmatic decisions (e.g., determining if a respondent has a psychological disorder or disability). Third, evaluating subtests entails evaluating their items, but the WIAT-4 authors provide little information about items in the technical manual. Although withholding this information from consumers became common practice in the mid-20th century, it is a lamentable practice because it precludes evaluation from disinterested scholars of interest (Buros 1977; Merton 1968).

3. Evaluation of the Wechsler Individual Achievement Test *Total Achievement Score*

The WIAT-4 authors state that the Total Achievement score "provides a measure of overall academic achievement in the areas of reading, math, and writing" (Breaux 2020, p. 113). Consequently, the first step in evaluating the validity of the Total Achievement score is understanding the meaning of the *overall academic achievement* (OAA) concept. Unfortunately, OAA is not a technical concept within either the psychology or education disciplines (i.e., it has no consistently shared meaning), and the WIAT-4 authors do not provide a definition. Thus, we need to explore the concept in more depth.

3.1. Meaning of Overall Academic Achievement

Psychologists have used OAA and similar terms for over a century, such as: general educational ability (Burt 1917), verbal-educational ability (Vernon 1950), scholastic achievement (Carroll 1943), schooling (French 1951), general academic intelligence (Dailey and Shaycoft 1961), and general academic achievement (Kaufman et al. 2012). With few exceptions, psychologists do not provide definitions or discuss the concepts' meanings except for stating it is distinct from, but related to, what Charles Spearman (1927) called *g*. In doing so, psychologists assume readers already understand the concepts, which means psychologists are likely employing ordinary language meanings. Although ordinary language concepts are not uncommon in psychology, they can be troublesome because they are often vague or ambiguous (Vygotsky 1987) which makes evaluating validity a particularly challenging endeavor (Haynes et al. 1995). Consequently, instead of understanding the meaning of OAA by working through a technical definition, we have to take a different tack. Specifically, we must (a) work through how psychologists employ the concepts of *overall*, *academic*, and *achievement* (i.e., conceptual analysis; Hacker and Stephan 2020); and then (b) reference those meanings to how the WIAT-4 authors discuss OAA and the procedures they use to measure it. Since the *overall* and *academic* terms modify *achievement*, we begin our conceptual work with achievement.

3.1.1. Meaning of *Achievement*

The unmodified *achievement* concept has a family of related meanings (Achievement 2021), but we will just focus on the two that psychologists seem to employ the most. One meaning is as a conative concept involving the desire to do things in such a way that they meet some standard (e.g., Heckhausen 1967). We manifest this need or motivation for achievement by doing things we believe will either avoid disapproval or attain approval

from ourselves or other persons (Crandall 1963). Psychologists have created different techniques and instruments to capture this form of achievement (e.g., projective testing, self-reports), but they all have in common coding respondents' behavior using some criteria other than correctness.

A second meaning of achievement is as the production of a particular outcome, either tangible (e.g., a loaf of bread) or intangible (e.g., goodwill from others). More specifically, it is an instantaneous and relatively durable effect of our behavior on situations (Vendler 1957). This meaning is intertwined with our knowledge and abilities to use knowledge, so is more of an intellective concept than conative (Reeve and Bonaccio 2011). As such, the techniques and instruments psychologists have created to capture this meaning of achievement commonly require coding behavior based on correctness (Guttman and Levy 1991).

Some psychologists claim that intellective achievement is a process more than an outcome (e.g., Bradford 2016; Coffman 1970), but this is likely better captured by the accomplishment concept. An *accomplishment* is a kind of goal-oriented process such that reaching the intended goal justifies employing the accomplishment term (Stokes 2008). That is, accomplishments are purposeful processes that culminate in something (i.e., an achievement). For example, if Pedro wrote a novel, it would be an accomplishment because writing a novel is something people have to commit to doing. The instant his novel is published, however, it is an achievement.

The distinction between achievement and accomplishment may appear trivial, but it is important (Varzi and Torrengo 2006). Achievements *can* be the culmination of a process designed to result in the achievements, but they can also result from a series of accidental or haphazard events. Accomplishments, however, cannot be accidental or haphazard. By definition, they are intentional culminations so depend on (a) knowledge about how to produce some achievement, and (b) the ability to employ the knowledge in such a way as to culminate in the particular achievement. Thus, Kiko responding to the item "3 + 2 = ?" correctly is an accomplishment only if she did so by employing her integer addition knowledge, but is an achievement irrespective of whether she employed her integer addition knowledge, guessed, or used some other process.

3.1.2. Meaning of *Achievement* in the Wechsler Individual Achievement Test

The scoring criteria for coding all responses to WIAT-4 items concern correctness, so we can deduce the instrument's authors employ the achievement concept in a way that is more consistent with the intellective meaning than the conative one. In addition, they employ the concept more consistent with an instantaneous outcome than a process. It is true that the authors discuss the mental processes they believe respondents should employ when answering items within a particular subtest, but this information was only used for item creation and designing procedures for WIAT-4 users to conduct a demand analysis (Breaux 2020, pp. 61–63). The actual mental processes respondents employ in their item responses are neither elicited or coded as part of the WIAT-4 administration nor used in the scoring procedures.

3.1.3. Meaning of *Academic* with Respect to *Achievement*

The unmodified achievement concept has a wide meaning and encompasses a variety of behaviors. As such, it is more a class of psychological attributes (i.e., umbrella concept) than a particular attribute. To limit the concept's boundaries, psychologists add a variety of modifying terms (e.g., athletic, occupational), but we only focus on the academic modifier. The *academic* concept has a few different meanings, but they are closely interwoven and all relate to school or education (Academic 2021). Thus, *academic achievements* are achievements that people manifest either in formal educational settings or result from abilities acquired from knowledge typically taught as part of formal education (Ebel and Frisbie 1991). This is still a very wide concept, including everything from alphabetic letter knowledge to

diagnosing a complex medical disorder correctly. Thus, psychologists typically take one of two tacks to further constrain the concept (Spinath 2012).

First, psychologists employ more domain-constraining modifiers (e.g., biochemistry achievement, nursing achievement). Psychologists typically do this when discussing achievements involving knowledge or abilities tied to particular curricula, so instruments designed to assess these achievements are also tied to curricula (e.g., curriculum-based assessments, licensing exams). Second, psychologists constrain the academic achievement concept to mean basic competencies typically acquired by members of a particular society or across multiple societies at certain ages. These competencies usually involve reading, writing, and using mathematics (Burt 1917; Mather and Abu-Hamour 2013). They are not tied to any particular curriculum, however, because psychologists create the instruments (a) to capture attributes that have some universality, and (b) for use with most or all societal members (Norenzayan and Heine 2005).

3.1.4. Meaning of *Academic Achievement* in the Wechsler Individual Achievement Test

The WIAT-4 authors do not discuss any particular curricula, but do discuss how differences in respondents' curriculum exposure can cause interpretational difficulties of some WIAT-4 scores (Breaux 2020, pp. 68, 72). Moreover, the Total Achievement score is comprised of scores from subtests in the reading, writing, and mathematics domains (see Table 1). Thus, we can infer that the WIAT-4 authors employ the academic achievement part of OAA to mean certain competencies members of American societies are expected to acquire.

3.1.5. Relation between Academic Achievement and Intelligence Instruments

If an instrument that captures academic achievement is not tied to any particular curriculum, captures somewhat universal abilities, and applies to most-or-all members of a society, then this naturally raises the question of how academic achievement instruments relate to intelligence instruments. Psychologists have a long history of discussing academic achievement and intelligence instruments as being distinct kinds (e.g., Matsumoto 2009). This is because psychologists have traditionally viewed academic achievement and intelligence as being distinct kinds of attributes (Anastasi 1984). Intelligence comprises a person's aptitude or potential to learn, while academic achievement is what a person has actually learned. The traditional view is flawed (Anastasi 1980; Wesman 1956). Support for this claim comes from the defining features of intelligence and intelligence instruments.

Intelligence is an ordinary language concept whose meaning has changed over time and geography (Goodey 2011; Spearman 1937). It entered the discipline of psychology in the 19th century by way of evolutionary biology (Danziger 1997). Biologists employed the concept as if it was a single attribute more or less synonymous with adaptive behavior or behavior flexibility. Psychologists tended to follow the biologists lead and employ the concept as if it was a single attribute, but not necessary one involving behavior flexibility/adaptation (cf. Bascom 1878; Taine 1872). Thus, there was ambiguity in the concept from the beginning.

Instead of reigning in the concept's meaning, however, psychologists in the early 20th century loosen it via their various idiosyncratic employments (e.g., Rugg 1921).[7] The concept eventually got so muddled that it became "a mere vocal sound, a word with so many meanings that finally it has none" (Spearman 1927, p. 14). One solution to this problem has been to re-define intelligence in such a way as to incorporate multiple existing meanings (e.g., Wechsler 1975). The major difficulty with this solution is that the resulting concepts are typically too vague to be measurable. A second solution is to invent new concepts that have a particular meaning and, often, a unique name (i.e., neologisms). Perhaps the best-known example is Spearman's invention of the *g* concept. Importantly, he did so with the intention of creating a technical concept amenable to scientific investigation, not to redefine intelligence (e.g., Spearman 1927, 1933, 1938). Thus, the major difficulty with this solution is that it does not address the ambiguity of the intelligence concept.

A third solution is to employ intelligence as an umbrella concept capturing a class of related attributes rather than one particular attribute (Howard 1993). This was how Spearman employed the concept (e.g., Spearman and Jones 1950), as did many of his protégés (e.g., Cattell 1987). This tradition continues today, with a recent conceptual study of intelligence concluding that intelligence "is a generic term, which encompasses a variety of constructs and concepts" (Reeve and Bonaccio 2011, p. 188). A major issue with this third solution is determining the criteria for an attribute to be included or excluded. Although psychologists have discussed multiple criteria, it appears that all intellective attributes share at least three major features (Burt 1944; Hacker 2013).

First, they involve our abilities to do something rather than mental states, dispositions, or attitudes. Second, these abilities involve acquiring or employing knowledge more than bodily movement (i.e., physical attributes), feelings/emotions (i.e., affective attributes), or motivation/volition (i.e., conative attributes). We discussed earlier that both features apply to academic achievement competencies as well. That is, psychologists tend to use the academic achievement concept to mean a class of abilities involving the employment of knowledge typically acquired in formal educational settings (Monroe et al. 1930).

Third, the abilities exist on a spectrum (Carroll 1993). When discussing intellective attributes, psychologists typically discuss this spectrum by referencing the breadth of tasks that elicit the attribute. At one end of the spectrum are *specific abilities* that people employ for a narrow set of tasks, while at the other end are *broad abilities* that people employ for a wide variety of tasks. In the context of academic achievement attributes, psychologists discuss the spectrum by referencing the specificity of a knowledge domain (Reeve and Bonaccio 2011). At one end of this spectrum is *domain-specific knowledge* that has a very circumscribed applicability (e.g., history of Leeds, England), while at the other end is *domain-general knowledge* that has much wider applicability (e.g., how to construct a valid argument).

Domain-specific knowledge and specific intellective abilities are not exchangeable concepts, but they are not unrelated either (and likewise for domain-general knowledge and broad intellective abilities). Instead, they represent differences in emphases (Reeve and Bonaccio 2011). Thus, it is better to think of the academic achievement and intelligence concepts as differing in degree more than in kind (Anastasi 1984; Cronbach 1990). That is, they are abilities that exist on a spectrum ranging from involving specific knowledge applicable to a very narrow range of tasks to those involving more general knowledge applicable to a broad array of tasks (Anastasi 1976; Carroll 1993; Schneider 2013).

Since intellective attributes all share some common features, it is not surprising that the multiplicity of intelligence instruments also shares a set of features (Guttman and Levy 1991). These instruments (a) contain items that elicit specific behavioral responses from examinees; (b) require examinees exert maximal effort in responding to items; and (c) provide guidelines for coding responses based on satisfying some logical, factual, or semantic rules (i.e., correctness). These features apply to academic achievement instruments as well (Thorndike and Thorndike-Christ 2010). Thus, irrespective of whether psychologists use the term intelligence or academic achievement in an instrument's name, the instrument measures (or potentially measures) the strength of one or more abilities a respondent has developed and is willing to demonstrate (Anastasi 1976, pp. 399–400).

3.1.6. Meaning of *Overall*

Overall is a somewhat ambiguous concept that can mean everything (i.e., end to end), operating over an entire range of things, or taking everything into consideration. The WIAT-4 authors provide some help narrowing the meaning because they use the term *general academic achievement* as a synonym for OAA (Breaux 2020, p. 42). Thus, we can infer that they believe the overall and general concepts are interchangeable. Unfortunately, *general* is not exceptionally clear in its meaning. In psychology, it has at least three meanings: breadth, depth, and summary (Beaujean 2015b; Spearman 1927).

As *breadth*, general concepts have more elements (i.e., broader) than more specific (i.e., narrow) concepts. In measurement models, this relation is often represented by a bi-factor structure whereby the *indicators* (i.e., recorded observations of phenomena, such as items or subtests) are specified to be the effects of (i.e., result from) both broader and narrower attributes operating more or less independently of each other (Holzinger et al. 1937). As *depth*, general concepts are at a higher level (i.e., super-ordinate) than more specific (i.e., sub-ordinate) concepts. In measurement models, this relation is often represented by a higher-order factor structure whereby (a) a set of indicators are specified to be the effects of multiple related attributes; and (b) those attributes are specified to be the unobserved (unmeasured) effects of more super-ordinate attributes.

As *summary*, general concepts and specific concepts both condense information with the difference being that general concepts condense over wider content than specific concepts. This relation can be represented by models with a formative-indicator structure (e.g., weighted average) or causal-indicator structure (Bollen and Bauldry 2011). Either way, the indicators are specified to influence the attributes rather than the attributes influencing the indicators. This entails that indicators define the attributes, so changing indicators can alter what instruments capture. This is not troublesome for instruments designed for pragmatic purposes (i.e., making diagnostic decisions) because authors create such instruments to produce scores that consistently predict some criteria external to the instrument (Burisch 1984). Having indicators define attributes is troublesome for measurement instruments, however, because it runs counter to the measurement process in science (Edwards 2011). Scientific measurement requires specifying an attribute's meaning before creating an instrument, which entails the meaning be invariant across indicators (Mari et al. 2015). Thus, it is unlikely that summary models are measurement models (Rhemtulla et al. 2015).

3.1.7. Meaning of *Overall (General)* in the Wechsler Individual Achievement Test

The WIAT-4 authors likely do not employ the overall/general concept to mean depth because they do not discuss OAA as influencing more narrow attributes (e.g., reading fluency). The authors are more equivocal about the breadth and summary meanings. On the one hand, they imply a summary meaning when they state the Total Achievement score provides "a midpoint for determining the examinee's relatively strong and weak areas of achievement" (Breaux 2020, p. 77). On the other hand, they imply a breadth meaning when they state the Total Achievement score provides an "overview of the examinee's overall achievement" and should be interpreted in a manner consistent with all the other WIAT-4 scores (e.g., report the score, confidence interval, and percentile rank; Breaux 2020, p. 77). Since the WIAT-4 authors are unclear about their meaning of overall/general, we will assume they mean having more breadth and, thus, consider whether OAA is a potentially measurable attribute.

3.2. Evidence for Functional Unity

Our brief conceptual analysis allows us to state that the WIAT-4 authors likely employ the OAA concept to mean a complex psychological attribute that involves employing abilities constitutive of reading, writing, and using mathematics. Reading, writing, and using mathematics all manifest in certain behaviors, which means OAA is observable, but may or may not be measurable. A necessary condition for OAA to be measurable is that the behaviors that constitute it have functional unity. We introduced the functional unity concept earlier but will expand upon it here.

A set of behaviors has functional unity when they are related in such a way that if any one of them changes, then the others "suffer the same fate" (Cattell 1956, p. 69). One line of evidence supporting functional unity comes from empirical investigations. Specifically, designing experiments to evaluate whether a set of behaviors "rise together, fall together, appear together, disappear together or, in general, covary together" (Horn 1972, p. 161). Empirical evidence is not sufficient, however, because behaviors could go together for reasons other than an attribute having functional unity (Coombs 1948). Thus,

in addition there needs to be a theory that provides a sound explanation for why the behaviors constitutive of a concept should hang together.

An example may clarify things. Pretend we have a battery with two subtests, both of which require respondents to listen and provide an oral response. One subtest contains items of the form "1 + 2 = ?", while the other contains items of the form "Do you believe that you often have to rushed to complete school work?". If we were to administer the battery to a set of elementary school students, it is not improbable that we would find that scores for the two subtests correlate at a level statistically different than zero (Lykken 1968). Although this corroborates the functional unity hypothesis, the unity is likely superficial because there is no theory explaining why behaviors across the subtests would go together. Instead, non-zero correlations among the subtests likely result from both subtests' items having a common administration medium, response modality, and requiring respondents to remember information.

If performance on all the integer addition items involve employing the same attribute or set of attributes, then it is possible that the unity of integer addition behavior may go beyond the superficial to a causal construct. This should manifest in particular relations among item performances across people on a single occasion as well as within the same people across multiple occasions (Horn 1963; Zimprich and Martin 2009). For example, if the integer addition items are arranged in order of increasing difficulty, then we would expect that for all respondents who correctly answered item p, then the probability of the same respondents answering items 1, 2, ... , $p - 1$ correctly is $\approx$1.00 (Loevinger 1947). Likewise, if we intervene with a particular student's integer addition skills, then not only should the student be able to answer item q ($q > p$) correctly, but also be able to answer items 1, 2, ... , $q - 1$ correctly as well. A possible explanation of this functional unity comes from the fact that mathematics is largely a graduated knowledge domain, so the ability to use more fundamental mathematics knowledge (e.g., adding integers without carrying) is usually necessary before being able to understand and use more advanced knowledge (e.g., adding integers with carrying).

3.2.1. Empirical Evidence for Functional Unity

The WIAT-4 technical manual provides two sources of empirical evidence concerning functional unity of OAA. The first is a study in which the WIAT-4 authors investigated the relation between Total Achievement score values across time (i.e., 12–87 days) for a subset of the norming sample (Breaux 2020, pp. 20–24). If OAA has functional unity, then we should observe relatively large correlations values among the Total Achievement scores across such a relatively short period of time. The correlation values are indeed large (i.e., .93–.95), which corroborates the hypothesis of OAA having functional unity.

The second source of evidence is the correlations among the WIAT-4 subtests for the norming sample. If OAA has functional unity, then we should observe relatively strong correlations among the subtests that comprise Total Achievement score. The WIAT-4 authors support functional unity by relying on visual inspection of the correlations (Breaux 2020, p. 29), but this is subject to the same cognitive biases as other visual inspection of data. A more robust approach is to subject the correlations to a factor analysis (Loehlin and Beaujean 2016a). Since the WIAT-4 authors do not provide any factor analytic results, we conducted our own.

Factor Analysis of Wechsler Individual Achievement Test Norming Data

Data for the factor analyses came from the WIAT-4 norming sample, which consists of 1832 participants aged between 4 and 50 years and was stratified to be consistent with the 2018 U.S. Census information. The sample includes 120 participants for each year from age 4 to 16 years, 120 participants for the combined age range of 17–19 years, 100 participants between the ages of 20 and 30 years, and 52 participants between the ages of 31 and 50. All data was collected between October of 2018 and February 2020—before American schools closed due to the COVID-19 pandemic.

For all factor analyses, we used the subtest correlation matrices provided in the WIAT-4 technical manual (Breaux 2020, pp. 31–34). The technical manual provides combined correlation matrices for the following age groups: 4–7 years (n = 480), 8–11 years (n = 480), 12–19 years (n = 720), and 20–50 years (n = 152). Some of the subtest scores are composite scores because they are comprised of two or more component scores. For the Listening Comprehension, Oral Expression, and Sentence Composition subtests, we included the composite score in the correlation matrix instead of the individual component scores.

For all factor analyses, we employed an unconstrained (i.e., "exploratory") model and used the entire correlation matrix rather than sets of particular subtests. We used the R statistical programming language (R Development Core Team 2017), particularly the *EFAtools* package (Steiner and Grieder 2020). Before initiating the factor extraction process, we subjected each correlation matrix to the Kaiser–Meyer–Olkin (KMO) test for sampling adequacy. KMO values were above .79 for each correlation matrix, so all matrices appear suitable for factor analysis.

To determine the number of factors to extract, we examined Kaiser's criterion method (Kaiser 1974), minimum average partial test (MAP; Velicer 1976), and parallel analysis (Horn 1965). The results are given in the right part of Table 2. The MAP test routinely suggested the presence of three factors, eigenvalues derived from the subtest correlation matrices ranged from 3 to 4, and parallel analysis suggested five factors for all but the older age-group, where it suggested three factors. To gain additional clarity about the number of factors to extract, we used statistical measures of the goodness of fit for models with 3–5 extracted factors. The statistical indices were used are: χ^2 goodness-of-fit test, Akaike information criterion (AIC), Bayesian information criterion (BIC), root mean square error of approximation (RMSEA) and comparative fit index (CFI). The χ^2 goodness-of-fit test indicted that none of the models fit the data well for any of the age groups. The other fit indices indicated more factors produced increasingly better fit, although the change from the three- to the four-factor solution was noticeably larger than from the four- to five-factor solution.

Table 2. Indices informing on number of factors to extract.

Number of Factors	χ^2	*df*	CFI	RMSEA (95% CI)	AIC	BIC	Eigen > 1	Parallel Analysis	MAP
				4–7 Years					
5	604.055	61	.967	0.14 (0.13–0.15)	482.055	227.454	3	5	3
4	837.031	74	.954	0.15 (0.14–0.16)	689.031	380.170			
3	1080.827	88	.940	0.15 (0.15–0.16)	904.827	537.533			
				8–11 Years					
5	284.022	100	.991	0.06 (0.05–0.07)	84.022	−333.356	4	5	3
4	468.924	116	.982	0.08 (0.07–0.09)	236.924	−247.236			
3	747.181	133	.969	0.10 (0.09–0.10)	481.181	−73.933			
				12–19 years					
5	185.215	86	.995	0.05 (0.04–0.06)	13.215	−345.731	3	5	3
4	279.639	101	.991	0.06 (0.05–0.07)	77.639	−343.914			
3	412.250	117	.986	0.07 (0.06–0.08)	178.250	−310.083			
				20–50 years					
5	458.797	86	.978	0.10 (0.09–0.10)	286.797	−72.148	4	3	3
4	594.135	101	.971	0.10 (0.09–0.11)	392.135	−29.417			
3	803.818	117	.960	0.11 (0.10–0.12)	569.818	81.485			

Note. AIC: Akaike information criterion; BIC: Bayesian information criterion, RMSEA: root mean square error of approximation, CFI: comparative fit index, MAP: Minimum average partial. χ^2 and χ^2-based fit indices (CFI, RMSEA, AIC, and BIC) were estimated used maximum likelihood extraction.

Given the ambiguity of the criteria for choosing the number of factors, we extracted 3–5 factors for each of the correlation matrices using the principal axis technique. We rotated the factors using a bi-factor rotation (Jennrich and Bentler 2011).[8] We did so because it

allows for a general factor (representing OAA) and multiple non-overlapping group factors that possibly represent more specific attributes. We conducted the bi-factor rotation using the procedures described by Loehlin and Beaujean (2016b) using 1000 random starting values and retaining the 10 best solutions. When the analysis returned multiple solutions, we retained the one with the lowest minimization value. When interpreting the loadings, we considered .3 to be a lower bound for a salient loading.

The results from our factor analysis indicate that the subtests that comprise Total Achievement score do tend to form a breadth factor (Tables A3, A5 and A7 in Appendix A). Across factor extractions and within each age group, all factor loadings on the general factor are above the salience criterion and are in same direction. At the same time, the factor loadings for some of the subtests appear to change noticeably across the age groups, especially for the oldest age group (20–50 years). For instance, in the solution with five specific factors, Essay Composition's general factor loading appears to drop substantially between the 12–19 and the 20–50 age group. This is currently just a hypothesis, however, because a rigorous evaluation of invariance is well beyond the scope of this article. Thus, we can state that there is some empirical evidence corroborating functional unity of OAA within an age group, but it is unknown if the unity exists across age groups.

3.2.2. Theoretical Evidence for Functional Unity

The technical manual contains no theoretical rationale for why the subtests that comprise OAA (as captured by the Total Achievement composite score) should hang together, much less a rationale for why some subtests might lose strength as indicators in adult respondents. Thus, we examined the intelligence and academic achievement literature for possible theories. One we believe is particularly useful is *triadic theory* (Cattell 1987; Cattell and Johnson 1986).[9]

In triadic theory, so-called crystallized intelligence (g_c) represents our cumulative knowledge across all knowledge domains. Triadic theory's investment aspect metaphorically explains g_c as resulting from the investments of our broader intellective attributes (e.g., memory, fluid intelligence), conative attributes (e.g., interests), and formal and informal educational opportunities. In school-age children, g_c often appears to be unitary across people, but this is not because g_c has functional unity. Instead, it is an artifact of strong developmental and situational constraints (e.g., similar interests, similar school curricula). Once the constraints weaken, g_c begins to differentiate (dissociate) into more specific attributes comprised of more specific knowledge (e.g., vocational, avocational).

To the extent that OAA and g_c are the same or strongly overlapping concepts, we would expect that the factor loadings for the subtests that comprise the Total Achievement score would weaken across age, especially in adulthood. This is because schooling is compulsory in the United States until the beginning of emerging adulthood (approximately 18 years of age). The fact that major differences in the WIAT-4 factor loadings are more or less confined to the oldest age group is consistent with predictions from the investment theory aspect of triadic theory. Of course, there could be other explanations that are just as consistent with the observed factor loadings. Until such explanations are put forth, however, we do not believe there is a theory-based justification for believing that OAA has functional unity. As such, it is not measurable and, thus, the Total Achievement score cannot have measurement validity.

4. Other Composite Scores

Step 2a in the WIAT-4 score interpretation guidance involves interpreting the other composite scores. We focus only on the composite scores in the domains of reading, writing, and mathematics because the WIAT-4 authors state that the fourth domain (i.e., oral language) is "not a core area of achievement" (Breaux 2020, p. 114).

4.1. Reading

The WIAT-4 authors created the reading domain subtests to align with the *simple view of reading* theory and its extensions (Hoover and Gough 1990; Kilpatrick 2015). The simple view of reading explains reading achievement as resulting from two conceptually independent mental attributes: word decoding/reading and oral language/linguistic comprehension. *Word decoding/reading* is the ability to apply knowledge of the relations between printed language and spoken language. It requires *cipher skills* (i.e., knowledge of letter–sound correspondences) and *word-specific knowledge* (i.e., applying cipher skills to particular words). *Oral language/linguistic comprehension* is the ability to apply knowledge of the oral language in which the words are written. Later extensions of the simple view of reading include contextual reading fluency as a bridge concept linking word decoding/reading and oral language/linguistic comprehension with reading comprehension. *Contextual reading fluency* is the speed at which we can accurately read connected text.

The WIAT-4 provides multiple subtests designed to capture word decoding/reading along with composite scores for cipher skills and word-specific knowledge (see Table 3). The three cipher skills composite scores are: Basic Reading, Decoding, and Phonological Processing. *Basic Reading* is "a composite score that closely aligns with the definition of basic reading skills specified by IDEA (2004) and many state guidelines for identifying specific learning disabilities" (Breaux 2020, p. 113).[10] The *Decoding* composite "provides an estimate of decontextualized phonic decoding and word reading skill" (Breaux 2020, p. 113), while *Phonological Processing* "measures phonemic proficiency and phonic decoding" (Breaux 2020, p. 114). The three composite scores are not independent, since the Pseudoword Decoding subtest is part of all three composites, while the Phonemic Proficiency and Word Reading subtests are both part of two composites. The WIAT-4 authors do not provide a justification for their rationale for having three strongly overlapping cipher skills composite scores.

Table 3. WIAT-4 Subtest and Composite Scores Aligned with the Simple View of Reading and its Extensions.

Reading Component	WIAT-4	
	Subtests	**Composites**
Word Decoding/Reading: Cipher Skills	Decoding Fluency	Basic Reading
	Phonemic Proficiency	Phonological Processing
	Pseudoword Decoding	Decoding
	Word Reading	
Word Decoding/Reading: Word-Specific Knowledge	Orthographic Choice [a]	Orthographic Processing
	Orthographic Fluency	Orthographic Processing Extended [a]
	Spelling [b]	
Oral Language/Linguistic Comprehension	Oral Language [c]	Listening Comprehension [c]
	Oral Expression [c]	Oral Expression [c]
Contextual Reading Fluency		Oral Reading Fluency
Reading Comprehension		Reading Comprehension

Adapted from Breaux (2020, p. 91). [a] Only available on Q-Interactive version of the instrument. [b] Part of writing domain. [c] Part of oral language domain.

The two composite scores capturing word-specific knowledge are *Orthographic Processing* and *Orthographic Processing Extended*. They both provide "an overall measure of orthographic processing, including the size of an examinee's orthographic lexicon and the quality of orthographic representations" (Breaux 2020, p. 114).[11] The difference between the scores is that the extended version includes one additional subtest that is only available on the Q-Interactive version of the instrument (Orthographic Choice). Both composite scores involve the Orthographic Fluency subtest as well as the Spelling subtest, the latter of which is part of the writing domain.

Since contextual reading fluency and reading comprehension are both captured by a single subtest, there are no composite scores for them. There is one composite score

capturing oral language/linguistic comprehension (Oral Language), which is comprised of two subtests in the oral language domain. As we noted earlier, however, the WIAT-4 authors do not include oral language as a core area of academic achievement (Breaux 2020, p. 114).

In addition to the theory-derived composite scores, there are two atheoretical composite scores in the reading domain: Reading Fluency and Reading. *Reading Fluency* "measures overall oral reading fluency skills" (Breaux 2020, p. 113). It consists of the Oral Reading Fluency, Orthographic Fluency, and Decoding Fluency subtests, although the latter is excluded in the composite for respondents not yet in third grade. The *Reading* composite score is comprised of the Word Reading and Reading Comprehension subtests, but the WIAT-4 authors are not explicit about what the composite score is designed to measure outside of stating it "balances word-level and text-level reading skills" (Breaux 2020, p. 112). According to the simple view of reading, word recognition and language comprehension represent distinct contributions to reading comprehension, so a change in students' reading decoding skills would not necessarily result in changing their reading comprehension. Thus, there is no reason to believe the Reading score captures an attribute with functional unity.

Empirical Evidence for Functional Unity of Reading Attributes

The WIAT-4 technical manual provides the same two sources of empirical evidence concerning functional unity of the behaviors comprising the reading attributes as it does OAA. The longitudinal study indicated relatively strong stability for all the composite scores, with all the correlation values greater than .90 (Breaux 2020, p. 22). This provides corroborating evidence for the hypothesis that the reading attributes represented by those scores have functional unity.

For the factor analysis, we employed the same data and data analysis procedures/programs as the OAA factor analysis except that we used promax rotation instead of bi-factor.[12] The results are given in Tables A2, A4 and A6. They indicate a messy structure for the reading subtests. The word decoding/reading subtests do not dissociate into cipher skills and word-specific knowledge, but instead all hang together along with the Oral Reading Fluency subtest. The oral language/linguistic comprehension subtests do comprise a different factor, but one with the Reading Comprehension and Math Problem Solving subtests—likely because these subtests all require significant language comprehension skills. In any case, the factor analysis does not provide strong evidence for functional unity of the attributes represented by the various reading composite scores. As such, it is difficult to make a strong argument that the composite scores have measurement validity.

4.2. Writing

The WIAT-4 authors created the writing subtests to be consistent with the simple view of writing and its extensions (Berninger and Winn 2006; Kim et al. 2018). In this theory, the working memory system (WM) coordinates the collective contributions of transcription skills, text generation/language skills, and self-regulation skills (i.e., executive functions) required for composition. *Transcription* involves both spelling and handwriting, while *text generation* involves the creation and organization of ideas as well as the language knowledge to transcribe the ideas into written text. All of these processes drain people's limited WM recourses, so the more writing skills people master (i.e., develop fluency) the more WM resources can be devoted to idea generation.

The WIAT-4 provides five subtests to capture the different aspects of writing, but their availability differs by grade (see Table 4). Alphabet Writing Fluency and Spelling capture transcription, while Sentence Composition and Essay Composition capture writing quality. Sentence Writing Fluency captures text writing fluency. The two oral language subtests (Listening Comprehension and Oral Expression) are the only subtests designed to capture text generation. The subtests constitute two writing composite scores: Writing Fluency and Written Expression (see Table 5). Both scores are troublesome.

Table 4. WIAT-4 Subtest Aligned with the Simple View of Writing and its Extensions.

Writing Component	Grades	WIAT-4 Subtests
Transcription	PK-4 K-12+	Alphabet Writing Fluency Spelling
Text Generation	PK-12+ PK-12+	Listening Comprehension [a] Oral Expression [a]
Text Writing Fluency	1–12+	Sentence Writing Fluency
Writing Quality	1–12 3–12+	Sentence Composition Essay Composition

Adapted from Breaux (2020, p. 95). [a] Part of oral language domain.

Table 5. WIAT-4 Writing Composite Scores.

Composite Score	Grades	Subtests
Writing Fluency	1–4	Alphabet Writing Fluency & Spelling
Written Expression	K–1	Spelling & Alphabet Writing Fluency
	2–3	Spelling & Sentence Composition
	4–12	Spelling, Sentence Composition, & Essay Composition

Adapted from Breaux (2020, pp. 112–13).

Empirical Evidence for Functional Unity of Writing Attributes

The *Writing Fluency* composite is comprised of the two transcription subtests, but the WIAT-4 authors do not discuss it as measuring transcription. Instead, they discuss it in term of a pragmatic purpose: capture developmental difficulties with both handwriting fluency and sentence-level text writing fluency for respondents in grades 1–4 (Breaux 2020, p. 113). Even if the WIAT-4 authors did make measurement claims about the score (i.e., represent transcription attribute), the claims would be difficult to support because of the low stability estimate for Writing Fluency is (i.e., .60; Breaux 2020, p. 23).

The WIAT-4 authors state that the *Written Expression* score "estimates overall written expression skills" (Breaux 2020, p. 112). This is neither an attribute within the simple view of writing nor an attribute the WIAT-4 authors discuss in any detail, so we have to infer its meaning based on subtest composition of the Written Expression score. The Written Expression score is comprised of Alphabet Writing Fluency, Essay Composition, Sentence Composition, and Spelling, but the particular subtests involved differ across respondent grade levels (see Table 5).

Across the entire norming sample, the stability estimate for the Written Expression score is .85 (Breaux 2020, p. 22). While this is relatively strong, there is little justification for believing the behaviors that constitute it have functional unity. Word, sentence, and text level writing build upon each other, but each level also requires unique skills. For instance, sentence-writing requires grammar knowledge not required in a spelling task, and text writing requires organizational skills not tapped by sentence-writing. As a result, writing tasks at different levels of language tend to not be highly associated with each other (Berninger et al. 1994). That was often the case in our factor analytic results (see Tables A1–A7). Spelling tended to load more with the decoding-oriented subtests in the reading domain, though often presented a small cross-loading with the writing measures. Although the Sentence and Essay Composition scores often loaded together, the loadings are noticeably weaker for the 20–50-year-old group than the other age groups.

4.3. Mathematics

All subtests in the mathematics domain are atheoretical. They were created to capture three areas in which people have mathematical difficulties: (a) math-fact fluency (i.e., recalling basic math facts quickly); (b) computation (i.e., understanding arithmetic operations and how they relate to each other and to apply computational procedures and strategies

fluently); and (c) math problem solving (i.e., applying knowledge to a problem for which the solution is not known, which is designed to enhance mathematical understanding and development).

There are two mathematics composite scores: Math Fluency and Mathematics. The *Math Fluency* composite provides "a measure of overall math fluency skills" in addition, subtraction, and multiplication (Breaux 2020, p. 113). It is comprised of between two to three Math Fluency subtests, depending on the respondents' grade level (see Table 6). The *Mathematics* composite "estimates overall mathematics skills in the domains of math problem solving and math computation" (Breaux 2020, p. 113), and is comprised of the Numerical Operations and Math Problem Solving subtests.

Table 6. WIAT-4 Mathematics Scores.

Area of Mathematics Difficulty	WIAT-4	
	Subtest Scores	**Composite Scores**
Math-fact fluency	Math Fluency–Addition	Math Fluency
	Math Fluency–Subtraction	
	Math Fluency–Multiplication [a]	
Computation	Numerical Operations	Mathematics
Math problem solving	Math Problem Solving	Mathematics

Adapted from Breaux (2020, pp. 112–13). [a] Available only for respondents in grade 3 or higher.

Empirical Evidence for Functional Unity of Mathematics Attributes

Across the entire norming sample, the stability estimates for both mathematics composite scores are greater than .90 (Breaux 2020, p. 22). Our factor analysis shows the Mathematics subtests do not hang together well. Across the different age groups, the Math Problem Solving subtest hangs together more with the oral language/reading comprehension subtests than any mathematics subtest. The Numerical Operations subtest joins this factor somewhat in the 12–19-year-old norming sample, and completely joins it in the in the 20–50-year-old-sample. Consequently, it is difficult to make an argument for interpreting the Mathematics composite score, much less believe that it has measurement validity. The Math Fluency subtests do appear to hang together well across all the age groups, which corroborates the hypothesis that the math fluency attribute has functional unity. As such, the Math Fluency composite could have measurement validity.

5. Evaluating the Equal-Interval Claim

Earlier we stated the WIAT-4 authors make a strong claim that some score values are on an equal-interval scale. The authors define an *equal-interval scale* as meaning "that a particular size of difference [i.e., interval] between two scores represents the same amount of difference in the skill [i.e., attribute] being measured regardless of where on the scale the scores fall" (Breaux 2020, p. 64). For example, if math fluency is measured on an equal-interval scale, then a change in Math Fluency score values from, say, 90 to 110 would represent the same change in the math fluency attribute as a score value change from 60 to 80. It is not uncommon for psychological instrument authors to claim that at least some of their score values have the equal-interval property (e.g., Kaufman et al. 2014, p. 91; Wechsler et al. 2014, pp. 14, 149) because it is necessary for many of the score interpretations that psychologists currently employ. For example, in the WIAT-4 the equal-interval property is necessary for interpretive steps 2a–4 as well as the two score analysis procedures the WIAT-4 authors suggest employing for identifying respondents with a specific learning disability/disorder (Breaux 2020, pp. 83–87).

Just as common as the equal-interval claim for psychological instrument scores is the lack of support for the claim.

To some extent this is understandable. Supporting the claim requires making the case that (a) the attribute of interest is a quantity, and (b) the score values that represent the attribute's manifestations preserve the attribute's quantitative features. Until the mid-20th

century it was largely believed that making such a case for psychological attributes was impossible (Michell 1999), and even now it is not straightforward how one goes about this (Markus and Borsboom 2013). We need not go into the detail here because the WIAT-4 authors neither provide support for their equal-interval claim, nor provide sufficient data in the technical manual for other psychologists to evaluate the claim empirically. Thus, we can only approach our evaluation of the equal-interval claim conceptually. We will do so for scores from two distinct, but typical, subtests: Numerical Operations and Math Fluency.

5.1. Numerical Operations

The *Numerical Operations* (NO) score "measures math computation skills" (Breaux 2020, p. 107) by capturing responses to items requiring mathematics computations ranging from naming numbers to basic calculus. By definition, if the math computation skills (MCS) attribute is a quantity, then it has the properties of equivalence, order, and additivity (Borsboom 2005; Hand 2004; Michell 1990). These are all technical concepts in measurement, but we can get by with their common-sense or intuitive meanings.

Equivalence roughly means that we can class any two people as either having distinguishable or indistinguishable forms of the attribute. If we can rank the distinguishable classes based on some feature of the attribute (e.g., amount, strength), then the attribute has *order*.[13] Having order means we can rank the equivalence classes, but tells us nothing about how much one class differs from another. It is only attributes with *additivity* that it makes sense to state whether the difference between any two classes is equivalent to the difference between any two other classes. For example, if MCS has additivity, then the difference between, say, the 10th ranked class and 20th ranked class is twice as much as the difference between the 15th ranked class in the 20th ranked class.

The WIAT-4 produces multiple value units for each score, but we focus first on the raw score unit. For NO raw score values to have equal-interval property, MCS needs to be a quantity and the NO raw score values need to represent MCS faithfully. That is, the NO raw score values need to represent MCS's equivalence, order, and additivity. If any one of these is not represented faithfully, then the NO raw score values cannot have equal intervals. We will assume MCS is a quantity and focus on NO representing its order property.

For the NO raw score values to represent the order of MCS faithfully, certain conditions must hold (Coombs et al. 1954). Specifically, the NO raw score values must be such that: (a) all respondents that are in the same MCS class (i.e., equivalent forms of MCS) have the same NO value, and all respondents that are in different MCS class (i.e., non-equivalent forms of MCS) have different NO values; (b) an order relation exists between respondents at each possible pair of NO values (e.g., respondents with a NO value of 100 have more MCS than respondents with a NO value of 99); and (c) there is consistency in the order relations (e.g., if respondents with a NO value of 100 have more MCS than those with a NO value of 99, and those with a NO value of 99 have more MCS than those with NO value of 98, then respondents with a NO value of 100 have more MCS than those with a NO value of 98). These conditions cannot be guaranteed to be true for the NO raw score values.[14]

The NO raw score is a behavior count consisting of the number of items a respondent correctly answered, and each item contributes the exact same to the raw score.[15] The items are not exchangeable, however, because they differ in content and difficulty. These features combine to allow for situations in which two respondents have the same NO score, yet answer different sets of items correctly and, potentially, have different MCS levels. For example, there are 495 ways to have a raw score of 4 on an instrument with 12 items.[16] Not all 495 of those patterns are possible, but if just one-fifth of them are, then that would still allow for 99 possible response patterns that produce a raw score of 4. The number of possible combinations of correct and incorrect responses expands rapidly as the number of items increases, and the NO subtest has over 50 items. Thus, it is highly probable that respondents with the same raw score have noticeably distinct response patterns. To the extent this is true, the structure of the NO raw score values is not guaranteed to represent

the order of MCS faithfully. As such, the NO raw scores could not represent the additivity of MCS, and thus, cannot comprise an equal-interval scale.

To some extent it is moot whether the raw score has validity because the WIAT-4 authors strongly discourage interpreting those score values—although for reasons other than we discussed (see Breaux 2020, p. 64). As an alternative, the WIAT-4 authors suggest interpreting one of the seven other units available for each score (i.e., standard, percentile rank, normal curve equivalent, stanine, age equivalent, grade equivalent, growth scale).[17] We will focus on the standard score unit because the WIAT-4 authors claim these values are on an equal-interval measurement scale (Breaux 2020, p. 64).

The WIAT-4 authors implicitly define a *standard score* using Equation (1) (Breaux 2020, p. 64).

$$\text{Standard} = \left(\frac{Raw - \overline{Raw}}{SD_{Raw}} \right) \times 15 + 100, \quad (1)$$

where Raw is the raw score for a particular respondent, $\overline{Raw}$ is the mean raw score in the selected norm group, and SD_{RAW} is the raw score standard deviation in the norm group.[18] An equivalent way of writing Equation (1) is in slope-intercept form, which is shown in Equation (2).

$$\text{Standard} = \frac{(100 \times SD_{Raw}) - (\overline{Raw} \times 15)}{SD_{Raw}} + Raw\left(\frac{15}{SD_{Raw}} \right). \quad (2)$$

Since $\overline{Raw}$ and SD_{RAW} are constants for a particular set of same-age or same-grade respondents, Equation (2) makes two things explicit. First, standard scores are just linear transformations of raw scores. As a linear transformation, the standard score conversion does not change anything about the raw score's structure, much less the structure of the represented attribute[19]. Instead, it just alters the meaning of the score values' origin (i.e., 0) and unit (i.e., 1). Thus, standard scores do not represent MCS any more faithfully than raw scores. If the raw scores values were not originally on an equal-interval scale, then the standard scores will not be one an equal-interval measurement scale either.

Second, standard score units are in standard deviations, so they are statistical units that represent variable dispersion within a group of respondents. They are not *measurement units*, which are particular manifestations of an attribute of interest used to represent other manifestations of the same attribute (Joint Committee for Guides in Metrology 2012). Score values expressed in standard deviations may have equal intervals on some statistical distribution, but it does not follow that the score values have equal intervals with respect to the attribute of interest. On the contrary, there is no reason to believe that changing some score unit to a standard deviation unit imbues the scores with any additional properties concerning the attribute of interest (Boring 1920).

An illustration will make this point more concrete. For kindergarten students in the fall of the academic year, average performing students (i.e., standard score of 100) can add single digits together, while students performing one standard deviation below the mean (i.e., standard score of 85) can identify numerals. The skill gap is starkly different from the same standard score difference for students in 12th grade. Average performing 12th grade students can solve algebraic equations and use geometry skills, while 12th grade students performing one standard deviation below the mean are likely struggling with fraction operations. Thus, even though the statistical unit-based scores are the exact same for both kindergarten and 12th grade students, the meaning of those scores with respect to MCS differs substantially.

5.2. Math Fluency

The *Math Fluency* composite provides "a measure of overall math fluency skills" (MFS) in addition, subtraction, and multiplication (Breaux 2020, p. 113). Each item in all three subtests consists of a single addition, subtraction, or multiplication problem that respondents solve correctly or incorrectly. There are two sets of items for the subtest, with

the set a particular respondent receives being based on the respondent's grade. Respondents complete as many problems as possible within 60 seconds.

Since fluency instruments are administered under strong time constraints, it is not uncommon to believe that raw scores from these instruments represent some attribute in an equal-interval unit (e.g., problems solved-correctly-per-minute). It is true that time is a base measurement unit for the physical sciences, so has equal-interval property. Nonetheless, dividing something by time does not necessarily put the resulting values in a base measurement unit (Boring 1920; Thomas 1942). This is because instruments designed to assess the speed of something and instruments designed capture speeded procedures are two different classes of instruments (Guttman and Levy 1991).

Instruments designed to assess speed are employed when time is part of the attribute's meaning. In psychology, the attribute is typically *response latency*, which is the time between the presentation of a stimulus (i.e., item) and the response. For example, if we are interested in measuring math fact retrieval speed, then we would present a math problem (e.g., "2 + 7 = ?") and immediately begin some a timing device that we would stop once the respondent provides the answer. Since scoring involves capturing the time it takes to respond rather than correctness, these instruments only contain items to which respondents are expected to answer correctly.

Instruments designed to capture speeded procedures consist of completing a set of items under strong time constraints. Typically, the constraints are so strong that respondents are not expected attempt all the items, and the non-response items are coded as being incorrect. Thus, responses are scored based on a correctness criterion rather than the time it takes to respond to any given item. This makes the raw score a count of the items correctly answered within in a certain period of time, which does not necessarily entail the values have an equal-interval unit (but see Johnson et al. 2019). The Math Fluency subtests belong to this class of instruments rather than the latency class. Thus, respondents who progress from, say, answering 50 problems per minute to answering 80 problems per minute do not necessarily have the same increase in MFS as respondents who progress from answering 270 problems per minute to answering 300 problems per minute—even though both changes involve 30 problems per minute.

To some extent, Math Fluency raw scores are irrelevant because the WIAT-4 authors provide no guidance for interpreting the values. Instead, they strongly suggest interpreting standard scores. As with the Numerical Operations subtest, however, transforming Math Fluency raw scores to standard scores does not give the score values additional properties with respect to representing the attribute of interest. Thus, if the Math Fluency raw score values do not have equal intervals, then there is no reason to believe that the Math Fluency standard score values will have equal intervals either.

6. Conclusions

The WIAT-4 is the latest iteration of a popular instrument designed to assess academic achievement in people across a wide variety of ages and grades. The WIAT-4 authors make two strong claims about the instrument: (a) the scores can be used for measurement purposes; and (b) some of the scores (i.e., standard scores) have values with equal intervals (Breaux 2020, pp. 1, 28, 64). Before psychologists adopt an instrument and interpret the scores in a manner consistent with the authors' claims, however, there should be sufficient evidence to support the claims (American Educational Research Association et al. 2014; International Test Commission 2001).

In this article, we evaluated the WIAT-4 authors' measurement claims (i.e., validity evidence) for the instrument's composite scores. Based only on the information provided in the WIAT-4 technical manual, we found the WIAT-4 authors did not provide sufficient evidence to support their measurement claims for the composite scores. First, many of the attribute concepts the scores ostensibly represent are ill defined in the technical manual (e.g., overall academic achievement) and it is unclear what attribute some of the scores are supposed to represent (e.g., Reading). As such, these scores' values cannot be measurement

values. Second, even for some of the attribute concepts with more clear meaning (e.g., cipher skills), the subtests that comprise the composite scores do not hang together in expected ways (i.e., do not appear to have functional unity). This makes it doubtful that the scores' values are measurement values.

There are a few attribute concepts the WIAT-4 authors discuss that have the potential for measurement (e.g., math fluency skills). For the scores to have measurement validity, however, the known properties of the attribute's manifestations need to be represented by the score values (Michell 1990; Joint Committee for Guides in Metrology 2012). Since the WIAT-4 authors claim that instrument's standard scores are on an equal-interval scale, this entails that (a) the attributes are quantities (i.e., manifestations have equivalence, order, and additivity); and (b) the relations among standard score values faithfully represent these relations among the attribute manifestations. The WIAT-4 authors provide no evidence in the technical manual to support their equal-interval claims, and our *prima facie* analysis of the claims was not favorable. As such, we highly doubt that the scores equal-interval properties. Thus, if the attributes the WIAT-4 capture are really quantities, then the scores that represent them are not doing so validity.

Practical Implications

The major practical implication of our evaluation concerns the appropriateness of the WIAT-4 authors' score interpretation guidance (Breaux 2020, pp. 77–79). First, step 1 in the interpretive guidance should be removed because the Total Achievement score should not be interpreted clinically. The score is supposed to represent overall academic achievement (OAA), but it is doubtful that OAA is even a clinically useful attribute concept, much less a unitary attribute. Based on the information provided in the technical manual, we cannot state that the Total Achievement score is anything more than a sum of items unique to the WIAT-4 instrument that the WIAT-4 authors believe are important.

Second, some of the composite scores may be useful for ranking students (step 2a in the interpretive guidance), but the evidence in the technical manual is insufficient to support the practice of interpreting quantitative differences in the composite scores (steps 2b–3) or subtest scores (step 4). Thus, any score comparisons should be limited to qualitative differences.

For example, pretend Zsa Zsa has an age-based Math Fluency standard score of 85 and a Reading Fluency standard score of 115. From this information, we can state her ability to conduct basic mathematics operations quickly is currently lower than the average ability of her same-age peers in the United States, while her oral reading ability for relatively simple English words is currently higher than her peers' average. It would be incorrect to interpret the 115–85 = 30-point difference between the scores because that the meaning of the 30-point difference differs across the score distributions. That is, even though numerically 115–85 = 90–60 = 130–100 = . . . , the meanings of the score differences with respect the represented attributes are not equivalent. Even qualitative interpretations of the differences in standard scores need to be done cautiously (Woodcock 1999). That is, interpreting the scores as indicating that Zsa Zsa's oral reading ability is "more developed" than her mathematics operations ability would not be warranted unless we had additional evidence (e.g., homework, motivation level; Shapiro 2011).

Although our evaluation is not supportive of the WIAT-4 authors' measurement validity claims, our evaluation is agnostic regarding whether psychologists should employ the instrument's scores for other purposes. Psychologists have a long history of employing instruments that produce scores that have utility (i.e., aid in making decisions) without measuring any attribute (e.g., Binet-Simon, Minnesota Multiphasic Personality Inventory; Berg 1959). Given the WIAT-4 authors' commendable revision of many scores in the reading and writing domains to align with strong theories in those areas, it is possible that those scores have utility for making decisions about respondents' academic achievement in those areas. The WIAT-4 authors do not provide the necessary information in the technical

manual to evaluate utility, however, so it will remain for future evaluations to determine whether WIAT-4 users should employ the scores for decision-making purposes.

On a final note, some readers of this article may believe that our evaluation of the WIAT-4 is out-of-sync with how psychologists currently think about validity and evaluate the validity of psychological instruments (e.g., Messick 1989). We acknowledge that the framework in which we evaluated the WIAT-4 is different from the received view of validity that permeates documents such as the American joint test standards (American Educational Research Association et al. 2014) or the European Federation of Psychologists' Associations model for instrument evaluation (Evers et al. 2013). We also acknowledge that the received view has been criticized extensively (e.g., Barrett 2018; Markus and Borsboom 2013). This criticism is not recent, however, but has a relatively long history in psychology. More than 40 years ago, Oscar Buros (1977) wrote, "If we make it our goal to measure rather than to differentiate, most of our methods of constructing tests, measuring repeatability, assessing validity, and interpreting test results will need to be drastically changed" (p. 12). It is our belief that our evaluation is fully in line with this needed drastic change.

Author Contributions: J.R.P. conducted the factor analysis. A.A.B. and J.R.P. conceived the article's ideas, discussed the results, and contributed to writing the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: AAB was supported by National Science Foundation Grant DRL1920730.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Data from this study came from the WIAT-4 Technical Manual.

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

Appendix A

Table A1. Oblique Factor Correlations Across Age Groups and Factor Models.

	Age 4 to 7					Age 8 to 11					Age 12 to 19					Age 20 to 50				
Five Factor Models																				
	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
Factor 1	1.00					1.00					1.00					1.00				
Factor 2	.66	1.00				.66	1.00				.72	1.00				.35	1.00			
Factor 3	.69	.57				.52	.58	1.00			.61	.62	1.00			.60	.13	1.00		
Factor 4	.73	.67	.63	1.00		.33	.40	.40	1.00		.64	.41	.41	1.00		.63	.45	.59	1.00	
Factor 5	.46	.36	.41	.50	1.00	.60	.62	.50	.29	1.00	.74	.76	.67	.44	1.00	.46	.34	.23	.58	1.00
SS Loadings [1]	4.85	2.39	1.45	1.43	1.16	4.94	2.70	3.09	.85	1.65	3.63	2.75	3.10	1.46	1.61	5.63	2.41	1.01	2.44	.88
Prop Tot Variance [1]	.29	.14	.09	.08	.07	.25	.13	.15	.04	.08	.19	.14	.16	.08	.08	.30	.13	.05	.13	.04
Prop Comm Var [1]	.43	.21	.13	.13	.10	.37	.20	.23	.06	.12	.29	.22	.25	.12	.13	.45	.20	.08	.20	.07
Four Factor Models																				
	F1	F2	F3	F4		F1	F2	F3	F4		F1	F2	F3	F4		F1	F2	F3	F4	
Factor 1	1.00					1.00					1.00					1.00				
Factor 2	.70	1.00				.65	1.00				.65	1.00				.37	1.00			
Factor 3	.58	.51	1.00			.51	.57	1.00			.56	.59	1.00			.63	.51	1.00		
Factor 4	.72	.65	.50	1.00		.61	.61	.52	1.00		.70	.77	.65	1.00		.55	.08	.40	1.00	
SS Loadings [1]	4.88	2.74	1.88	1.30		4.97	2.79	3.17	1.69		4.37	2.53	3.01	2.26		5.54	3.04	2.57	.82	
Prop Tot Variance [1]	.29	.16	.11	.08		.25	.14	.16	.08		.23	.13	.16	.12		.29	.16	.14	.04	
Prop Comm Var [1]	.45	.25	.17	.12		.39	.22	.25	.13		.36	.21	.25	.19		.46	.25	.21	.07	
Three Factor Models																				
	F1	F2	F3			F1	F2	F3			F1	F2	F3			F1	F2	F3		
Factor 1	1.00					1.00					1.00					1.00				
Factor 2	.71	1.00				.70	1.00				.60	1.00				.41	1.00			
Factor 3	.67	.61	1.00			.55	.58	1.00			.72	.64	1.00			.63	.55	1.00		
SS Loadings [1]	5.08	2.88	2.40			5.59	3.01	3.30			4.87	3.77	3.10			5.53	3.11	2.77		
Prop Tot Variance [1]	.30	.17	.14			.28	.15	.17			.26	.20	.16			.29	.16	.15		
Prop Comm Var [1]	.49	.28	.23			.47	.25	.28			.41	.32	.26			.48	.27	.24		

Note. Variance proportions may not appear in descending order. They were re-ordered for comparison to the tables of factor loadings; [1]—For oblique rotations, we report variance accounted via the rotated factor loadings, accounting for correlations between factors.

Table A2. Five Factor Oblique Solution Across Age Groups.

	Factor 1				Factor 2				Factor 3				Factor 4				Factor 5				Communalities [1]			
	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50
WR	**.98**	**.77**	**.76**	**.99**	−.04	.21	.24	.09	.14	−.13	−.09	−.14	−.06	−.12	.09	−.12	−.12	.14	−.08	−.06	.92	.83	.80	.79
RC	**.36**	.14	.02	.24	**.36**	**.67**	**.62**	**.41**	.10	−.08	−.05	.28	−.07	.09	.07	−.10	.08	.03	.13	−.02	.58	.61	.55	.55
PD	**.72**	**.86**	**1.07**	**.87**	.15	−.02	−.17	−.01	−.01	−.05	.03	−.04	−.09	−.06	−.03	−.04	.10	.07	−.08	.07	.73	.71	.82	.81
OF	**.99**	**.76**	.23	**.79**	.01	.19	.26	−.11	−.14	.08	.05	.11	.05	.04	**.55**	−.02	.02	−.14	.01	.10	.89	.77	.83	.83
DF	-	**1.02**	**.60**	**1.05**	-	−.13	−.10	−.19	-	.10	.11	-.08	-	.05	**.44**	.13	-	−.15	−.09	−.19	-	.86	.80	.81
ORF	**.94**	**.68**	.12	**.62**	−.01	.11	−.01	.10	−.29	.15	.07	.19	.17	.08	**.53**	−.02	−.04	−.16	.19	−.09	.65	.63	.59	.60
SP	**.56**	**.68**	**.54**	**.85**	−.23	−.10	.01	.01	.27	−.02	−.01	−.14	−.02	−.01	.12	.03	**.30**	**.41**	**.32**	.21	.74	.81	.77	.77
SC	.26	−.04	.13	.13	**.38**	.28	.12	.06	−.04	.03	−.05	.10	−.05	.14	−.05	−.06	.28	**.57**	**.60**	**.53**	.57	.68	.56	.56
EC	-	.17	−.07	−.13	-	−.25	−.12	.04	-	.05	-.02	.29	-	.06	.12	.04	-	**.75**	**.80**	**.51**	-	.57	.49	.50
AWF	.26	−.05	-	-	−.16	.01	-	-	**.32**	−.02	-	-	.17	**.75**	-	-	**.33**	.17	-	-	.31	.63	-	-
SWF	−.05	.03	−.05	−.08	.05	.19	.16	−.24	−.09	.23	.26	**.50**	.15	.30	.22	.18	**.74**	−.11	−.04	.13	.51	.30	.20	.20
MPS	−.11	−.02	.15	−.02	**.49**	**.59**	**.57**	**.91**	**.43**	**.36**	**.33**	−.20	.14	−.19	−.16	.18	−.15	.10	.02	.01	.69	.71	.77	.77
NO	.01	−.06	−.05	−.09	.00	.19	**.31**	**.92**	**.82**	**.59**	**.43**	**−.31**	.08	−.02	−.13	.21	−.04	.24	.29	.03	.63	.68	.68	.68
MFA	−.06	−.01	.05	.07	−.03	−.07	−.09	.11	.04	**.88**	**.96**	.18	**.76**	−.01	.05	**.81**	.17	−.02	−.10	−.12	.79	.68	.81	.80
MFS	.07	−.02	.02	.00	.12	−.01	.01	−.06	.12	**.94**	**.93**	**.30**	**.66**	−.05	.05	**.83**	.05	.01	−.08	.10	.75	.83	.84	.85
MFM	.09	.08	−.05	−.03	-	−.14	−.05	.10	-	**.84**	**.85**	-.02	-	.09	.07	**.82**	-	.01	.11	.00		.73	.80	.80
LC	−.05	.06	−.07	.16	**.89**	**.91**	**1.00**	**.32**	−.08	−.09	−.07	**.38**	.03	−.06	.08	−.08	−.11	−.21	−.21	.13	.58	.60	.64	.66
OE	−.03	−.02	.01	**.33**	**.79**	**.76**	**.67**	**.35**	−.06	−.02	−.04	.18	.02	.13	.10	−.08	.12	−.03	.09	.04	.66	.60	.59	.60
PP	.26	**.46**	**.68**	**.48**	**.34**	.18	.17	.12	.22	−.09	−.02	.26	−.13	.04	−.12	.11	.16	.21	−.01	−.06	.63	.50	.53	.56
OC	**.81**	**.58**	**.35**	**.64**	−.01	−.04	.03	.09	.13	−.02	.01	−.14	.09	−.06	.09	.16	−.23	.26	**.31**	.12	.65	.51	.51	.73

Note. Factor loadings greater than 1 suggest Heywood cases. [1]—Communalities represent the final estimates from the unrotated solution; WR—Word Reading; RC—Reading Comprehension; PD—Pseudoword Decoding; OF—Orthographic Fluency; DF—Decoding Fluency; ORF—Oral Reading Fluency; SP—Spelling; SC—Sentence Composition; EC—Essay Composition; AWF—Alphabet Writing Fluency; SWF—Sentence Writing Fluency; MPS—Math Problem Solving; NO—Numerical Operations; MFA—Math Fluency Addition; MFS—Math Fluency Subtraction; MFM—Math Fluency Multiplication; LC—Listening Comprehension; OE—Oral Expression; PP—Phonemic Proficiency; OC—Orthographic Choice; Columns represent age groups, 4 to 7, 8 to 11, 12 to 19, 20 to 50. **Bold** coefficients are greater than .30.

Table A3. Bifactor Solution with Five Specific Factors Across Age Groups.

	General Factor				Factor 1				Factor 2				Factor 3				Factor 4				Factor 5				Communalities			
	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50
WR	**.90**	**.85**	**−.87**	**.88**	−.25	−.09	−.05	.03	−.07	.01	−.03	.02	−.01	−.14	.13	−.13	.08	−.23	.05	−.04	.14	.12	.14	.14	.91	.82	.79	.81
RC	**.72**	**.69**	**−.66**	**.64**	−.07	.07	.01	.03	.22	**.34**	**.34**	**−.37**	.03	.00	.00	.01	.07	−.06	−.01	.22	−.03	−.12	−.03	−.04	.58	.61	.55	.59
PD	**.87**	**.72**	**−.84**	**.87**	−.01	−.14	.02	.18	.00	−.05	−.29	.00	−.25	−.04	.09	−.07	.06	**−.58**	−.03	−.08	.10	.05	.11	.01	.83	.88	.81	.80
OF	**.85**	**.84**	**−.83**	**.83**	**−.55**	**−.35**	**−.38**	**−.35**	.00	.02	.03	.03	.00	.01	−.02	−.07	.03	−.01	.03	−.07	−.05	.02	.03	−.01	1.02	.83	.83	.81
DF	-	**.77**	**−.80**	**.81**	-	**−.38**	**−.34**	−.03	-	−.11	−.24	.06	-	.03	.00	.03	-	**−.30**	−.03	−.26	-	.00	.07	.22	-	.84	.81	.77
ORF	**.67**	**.75**	**−.66**	**.73**	**−.43**	**−.36**	**−.38**	**−.31**	−.03	−.01	−.07	−.19	−.10	.06	−.05	−.02	−.09	.02	.02	.01	.03	−.02	−.09	.12	.65	.69	.60	.68
SP	**.84**	**.86**	**−.86**	**.90**	.00	−.01	−.04	.02	−.22	−.22	−.05	.13	−.02	−.05	.02	.01	−.01	−.15	.18	.03	−.06	.04	−.07	−.04	.76	.81	.77	.83
SC	**.72**	**.72**	**−.67**	**.50**	.05	**.34**	.11	−.06	.20	.03	.13	.06	−.13	.08	−.02	−.01	.03	.01	.02	.24	−.10	−.15	−.29	**−.35**	.58	.66	.56	.43
EC	-	**.63**	**−.56**	**.38**	-	**.31**	−.01	.06	-	**−.38**	.04	−.08	-	.00	−.06	.09	-	.07	.05	.20	-	−.03	**−.43**	**−.45**	-	.65	.50	.41
AWF	**.44**	**.38**	-	-	−.01	.05	-	-	−.10	−.01	-	-	.25	.11	-	-	−.13	.02	-	-	−.25	**−.72**	-	-	.34	.68	-	-
SWF	**.51**	**.37**	**−.38**	.19	.00	−.08	−.15	−.21	.01	.13	.06	−.23	−.02	.23	−.19	.13	−.14	.04	−.06	−.10	**−.58**	−.27	.00	−.19	.61	.29	.20	.20
MPS	**.68**	**.70**	**−.76**	**.56**	.01	.15	.18	−.02	**.36**	.29	**.30**	−.03	**.31**	**.30**	−.25	**.39**	−.09	.02	−.07	**.60**	.05	.11	.01	.01	.70	.70	.77	.82
NO	**.64**	**.65**	**−.67**	**.45**	.11	.16	.16	.06	.03	.06	.24	.05	**.45**	**.48**	**−.35**	**.42**	−.07	.03	.02	**.60**	.03	−.03	−.14	.00	.63	.68	.68	.75
MFA	**.60**	**.44**	**−.59**	**.51**	.01	−.02	−.03	.04	−.03	.00	−.06	−.09	.05	**.70**	**−.67**	**.76**	**−.77**	−.03	−.02	.00	−.04	−.02	.01	.02	.95	.69	.80	.84
MFS	**.64**	**.53**	**−.64**	**.53**	−.03	.01	−.02	.01	.07	.02	.01	−.06	.15	**.74**	**−.67**	**.75**	**−.50**	−.02	.02	−.02	−.02	.00	.03	−.15	.69	.84	.85	.86
MFM	-	**.53**	**−.63**	**.34**	-	−.08	−.02	−.08	-	−.08	.01	.14	-	**.65**	**−.63**	**.78**	-	.07	.02	.07	-	−.08	−.08	−.01	-	.73	.80	.75
LC	**.47**	**.58**	**−.59**	**.63**	.02	−.02	−.01	−.07	**.59**	**.50**	**.53**	**−.34**	.01	−.04	.03	.02	.00	.07	.02	.23	.02	.01	.19	−.15	.58	.60	.66	.59
OE	**.62**	**.63**	**−.68**	**.67**	.03	.06	−.01	−.10	**.52**	**.41**	**.35**	−.22	.00	.05	.01	.02	.01	.09	−.10	.22	−.11	−.15	−.03	−.03	.66	.60	.60	.55
PP	**.77**	**.65**	**−.72**	**.73**	.10	.09	.13	.13	.18	.07	−.04	**−.30**	−.02	−.03	.08	.12	.10	**−.40**	−.12	−.02	−.01	−.08	.04	.01	.64	.60	.56	.66
OC	**.74**	**.68**	**−.70**	**.74**	−.24	−.05	.00	.00	−.03	−.15	.00	.13	.03	−.07	−.01	.15	−.05	−.09	**.49**	.06	.19	.09	−.04	.00	.64	.51	.73	.59
ω / ω.h	.95	.89	.92	.85	.01	.00	.00	.00	.02	.00	.01	.01	.00	.05	.04	.07	.01	.01	.00	.02	.00	.01	.00	.01				
H	.96	.95	.96	.96	.45	.44	.37	.28	.56	.51	.50	.38	.35	.77	.72	.82	.65	.48	.28	.60	.40	.56	.30	.36				
SS Loadings	8.31	8.83	9.28	8.13	.64	.71	.52	.35	.94	.89	.84	.56	.48	1.89	1.54	2.16	.92	.70	.32	1.07	.50	.70	.38	.49				
Prop Tot Variance	.49	.44	.49	.43	.04	.04	.03	.02	.06	.04	.04	.03	.03	.09	.08	.11	.05	.03	.02	.06	.03	.04	.02	.03				
Prop Comm Var	.71	.64	.72	.64	.05	.05	.04	.03	.08	.06	.07	.04	.04	.14	.12	.17	.08	.05	.02	.08	.04	.05	.03	.04				

Note. WR—Word Reading; RC—Reading Comprehension; PD—Pseudoword Decoding; OF—Orthographic Fluency; DF—Decoding Fluency; ORF—Oral Reading Fluency; SP—Spelling; SC—Sentence Composition; EC—Essay Composition; AWF—Alphabet Writing Fluency; SWF—Sentence Writing Fluency; MPS—Math Problem Solving; NO—Numerical Operations; MFA—Math Fluency Addition; MFS—Math Fluency Subtraction; MFM—Math Fluency Multiplication; LC—Listening Comprehension; OE—Oral Expression; PP—Phonemic Proficiency; OC—Orthographic Choice; Columns represent age groups, 4 to 7, 8 to 11, 12 to 19, 20 to 50. **Bold** coefficients are greater than .30.

Table A4. Four Factor Oblique Solution Across Age Groups.

	Factor 1				Factor 2				Factor 3				Factor 4				Communalities [1]			
	4–7	**8–11**	**12–19**	**20–50**	**4–7**	**8–11**	**12–19**	**20–50**	**4–7**	**8–11**	**12–19**	**20–50**	**4–7**	**8–11**	**12–19**	**20–50**	**4–7**	**8–11**	**12–19**	**20–50**
WR	**1.03**	**.80**	**.68**	**.98**	.01	.16	.23	.04	−.03	−.16	−.12	−.12	−.09	.11	.13	−.13	.92	.82	.77	.81
RC	**.36**	.14	.08	.21	**.41**	**.69**	**.59**	**.46**	−.05	−.06	−.03	−.09	.10	.05	.14	.25	.58	.62	.55	.55
PD	**.71**	**.88**	**.76**	**.87**	.15	−.04	−.11	.06	−.11	−.06	−.03	−.05	.12	.04	.23	−.06	.73	.72	.71	.74
OF	**.96**	**.76**	**.76**	**.78**	−.04	.20	.23	.02	.01	.10	.10	−.03	.00	−.14	−.14	.09	.87	.77	.78	.69
DF	-	**1.00**	**.98**	**1.04**	-	−.10	−.08	−.35	-	.13	.12	.15	-	−.15	−.15	−.07	-	.84	.82	.78
ORF	**.84**	**.66**	**.66**	**.59**	−.11	.14	−.01	.08	.07	.18	.13	.00	−.07	−.14	−.02	.19	.59	.62	.52	.58
SP	**.60**	**.69**	**.54**	**.85**	−.17	−.10	−.03	.17	.08	−.02	−.02	.01	**.39**	**.39**	**.45**	−.14	.73	.81	.78	.80
SC	.22	−.06	.05	.16	**.37**	**.33**	.06	**.48**	−.07	.03	−.04	−.09	**.30**	**.62**	**.69**	.03	.57	.70	.56	.32
EC	-	.17	.10	−.08	-	−.23	−.12	**.47**	-	.04	.05	.01	-	**.76**	**.61**	.16	-	.58	.40	.26
AWF	−.01	−.08	-	-	−.10	.24	-	-	**.31**	.16	-	-	**.39**	.22	-	-	.29	.22	-	-
SWF	−.15	−.01	.20	−.10	−.01	.29	.15	-.03	.17	**.30**	.28	.17	**.68**	−.04	−.14	**.43**	.45	.24	.20	.17
MPS	.08	.05	−.08	−.02	**.65**	**.48**	**.52**	**.81**	.28	**.31**	.28	.21	−.11	.07	.23	−.15	.66	.62	.75	.78
NO	.13	−.04	−.19	−.09	.24	.18	.26	**.83**	**.32**	**.57**	**.40**	.23	.09	.23	**.41**	−.27	.43	.67	.68	.76
MFA	.02	−.01	.09	.06	−.08	−.07	-.08	.00	**.80**	**.88**	**.92**	**.85**	.16	−.03	−.05	.18	.74	.68	.80	.83
MFS	.06	−.01	.06	.01	.10	−.02	.01	.03	**.75**	**.92**	**.90**	**.84**	.04	.00	−.03	.25	.76	.82	.84	.85
MFM	-	.07	.02	−.02	-	−.10	−.06	.02	-	**.86**	**.84**	**.85**	-	.02	.12	−.02	-	.73	.80	.73
LC	−.10	.08	.01	.12	**.91**	**.85**	**.97**	**.51**	−.02	−.09	−.06	−.09	−.15	−.20	−.20	**.34**	.56	.57	.65	.61
OE	−.08	−.04	.10	**.31**	**.81**	**.80**	**.63**	**.43**	−.02	.01	−.02	−.07	.09	.00	.09	.17	.65	.61	.59	.55
PP	.28	**.46**	**.39**	**.46**	**.44**	.19	.17	.13	−.08	−.08	−.07	.13	.22	.20	.26	.23	.63	.50	.48	.58
OC	**.86**	**.60**	**.37**	**.64**	.03	−.06	−.01	.16	.13	−.04	.00	.15	−.24	.24	**.41**	−.14	.66	.50	.51	.58

Note. Factor loadings greater than 1 suggest Heywood cases. [1]—Communalities represent the final estimates from the unrotated solution; WR—Word Reading; RC—Reading Comprehension; PD—Pseudoword Decoding; OF—Orthographic Fluency; DF—Decoding Fluency; ORF—Oral Reading Fluency; SP—Spelling; SC—Sentence Composition; EC—Essay Composition; AWF—Alphabet Writing Fluency; SWF—Sentence Writing Fluency; MPS—Math Problem Solving; NO—Numerical Operations; MFA—Math Fluency Addition; MFS—Math Fluency Subtraction; MFM—Math Fluency Multiplication; LC—Listening Comprehension; OE—Oral Expression; PP—Phonemic Proficiency; OC—Orthographic Choice; Columns represent age groups, 4 to 7, 8 to 11, 12 to 19, 20 to 50. **Bold** coefficients are greater than .30 **Bold** coefficients are greater than .30.

Table A5. Bifactor Solution with Four Specific Factors Across Age Groups.

	General				Factor 1				Factor 2				Factor 3				Factor 4				Communalities			
	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50	4–7	8–11	12–19	20–50
WR	**.94**	**.88**	**−.87**	**.88**	.06	−.01	−.04	.03	.11	.14	.05	.05	.07	−.16	.14	.14	.16	−.14	.09	.16	.92	.83	.80	.82
RC	**.72**	**.69**	**−.66**	**.64**	−.23	**.35**	**.34**	**−.35**	.06	−.11	−.01	−.21	−.02	−.02	.01	−.01	−.06	.04	−.05	−.03	.58	.61	.55	.58
PD	**.84**	**.79**	**−.84**	**.85**	−.07	−.13	−.30	.03	.11	.10	−.02	.06	.10	−.11	.09	.07	−.02	−.23	.09	.04	.73	.71	.82	.73
OF	**.91**	**.83**	**−.83**	**.82**	.04	.04	.03	−.05	.02	−.01	**.39**	.11	.22	.03	−.01	.07	.08	**−.30**	.01	−.01	.89	.77	.83	.69
DF	-	**.81**	**−.79**	**.80**	-	−.17	−.23	.07	-	.01	**.33**	.28	-	.01	.00	−.01	-	**−.42**	.07	.25	-	.86	.80	.78
ORF	**.73**	**.73**	**−.66**	**.74**	.06	.01	−.07	−.22	−.06	−.05	**.38**	.05	.30	.09	−.05	.02	.12	−.30	−.10	.08	.65	.63	.59	.60
SP	**.83**	**.87**	**−.86**	**.90**	.18	−.21	−.07	.15	.00	.05	.04	−.01	−.05	−.07	.03	.00	−.14	.00	−.14	−.04	.74	.81	.77	.83
SC	**.69**	**.72**	**−.66**	**.51**	−.24	.06	.13	.05	.04	−.14	−.11	−.20	.04	.06	−.02	.01	−.19	**.36**	**−.30**	**−.37**	.57	.68	.56	.44
EC	-	**.61**	**−.54**	**.38**	-	−.29	.04	−.07	-	−.03	.02	−.19	-	.02	−.07	−.09	-	**.34**	**−.43**	**−.42**	-	.57	.49	.37
AWF	**.44**	**.39**	-	-	.10	.00	-	-	−.20	**−.68**	-	-	-.16	.10	-	-	-.20	.09	-	-	.31	.63	-	-
SWF	**.49**	**.37**	**−.37**	.20	−.02	.14	.07	−.25	−.19	−.29	.15	.13	.05	.24	−.19	−.13	**−.48**	−.06	.01	−.18	.51	.30	.20	.17
MPS	**.68**	**.70**	**−.76**	**.58**	**−.36**	**.31**	**.31**	−.04	−.15	.12	−.18	**−.58**	−.27	.30	−.24	**−.38**	.07	.13	.01	.00	.69	.71	.77	.81
NO	**.63**	**.65**	**−.67**	**.47**	−.04	.08	.24	.05	−.15	−.02	−.16	**−.61**	**−.46**	**.47**	**−.35**	**−.41**	.02	.18	−.16	−.01	.63	.68	.68	.77
MFA	**.59**	**.45**	**−.59**	**.52**	.02	−.01	−.05	−.06	**−.66**	−.03	.03	−.01	.00	**.69**	**−.67**	**−.76**	−.05	−.01	.02	.05	.79	.68	.80	.84
MFS	**.64**	**.54**	**−.64**	**.53**	−.08	.02	.01	−.04	**−.57**	.00	.02	.02	−.05	**.73**	**−.66**	**−.75**	.01	.02	.02	−.13	.75	.83	.84	.86
MFM	-	**.53**	**−.63**	**.35**	-	−.06	.01	.14	-	-.10	.02	-.07	-	**.66**	**−.63**	**−.76**	-	−.04	−.09	−.01	-	.73	.80	.73
LC	**.46**	**.56**	**−.60**	**.64**	**−.61**	**.53**	**.51**	**−.35**	−.01	.00	.01	−.20	−.02	−.02	.04	−.02	.00	−.04	.14	−.15	.58	.60	.64	.60
OE	**.60**	**.61**	**−.68**	**.68**	**−.54**	**.44**	**.36**	−.24	−.02	−.16	.01	−.19	−.02	.05	.00	−.01	−.14	.06	−.02	−.04	.66	.60	.59	.55
PP	**.74**	**.70**	**−.71**	**.73**	−.23	.01	−.02	−.24	.09	−.02	−.12	.01	−.10	−.09	.08	−.12	−.12	−.01	.05	.03	.63	.50	.53	.60
OC	**.77**	**.69**	**−.70**	**.74**	.03	−.14	−.02	.14	−.03	.09	.03	−.05	.04	−.07	.00	−.14	.23	−.04	−.15	.00	.65	.51	.51	.59
ω/ω.h	.92	.91	.92	.87	.03	.01	.01	.01	.02	.01	.00	.02	.00	.05	.04	.07	.00	.00	.00	.00				
H	.96	.96	.96	.96	.58	.53	.50	.38	.59	.52	.37	.59	.36	.77	.71	.82	.35	.46	.32	.36				
SS Loadings	8.37	9.00	9.26	8.20	1.02	.93	.84	.56	.93	.67	.52	1.02	.48	1.88	1.53	2.11	.47	.74	.41	.49				
Prop Tot Variance	.49	.45	.49	.43	.06	.05	.04	.03	.05	.03	.03	.05	.03	.09	.08	.11	.03	.04	.02	.03				
Prop Comm Var	.74	.68	.74	.66	.09	.07	.07	.04	.08	.05	.04	.08	.04	.14	.12	.17	.04	.06	.03	.04				

Note. WR—Word Reading; RC—Reading Comprehension; PD—Pseudoword Decoding; OF—Orthographic Fluency; DF—Decoding Fluency; ORF—Oral Reading Fluency; SP—Spelling; SC—Sentence Composition; EC—Essay Composition; AWF—Alphabet Writing Fluency; SWF—Sentence Writing Fluency; MPS—Math Problem Solving; NO—Numerical Operations; MFA—Math Fluency Addition; MFS—Math Fluency Subtraction; MFM—Math Fluency Multiplication; LC—Listening Comprehension; OE—Oral Expression; PP—Phonemic Proficiency; OC—Orthographic Choice; Columns represent age groups, 4 to 7, 8 to 11, 12 to 19, 20 to 50. **Bold** coefficients are greater than .30.

Table A6. Three Factor Oblique Solution Across Age Groups.

	Factor 1				Factor 2				Factor 3				Communalities [1]			
	4–7	**8–11**	**12–19**	**20–50**	**4–7**	**8–11**	**12–19**	**20–50**	**4–7**	**8–11**	**12–19**	**20–50**	**4–7**	**8–11**	**12–19**	**20–50**
WR	**1.00**	**.88**	**.73**	**.89**	−.01	.16	.29	.05	−.05	−.16	−.13	−.09	.91	.82	.77	.79
RC	**.40**	.14	.08	.31	**.45**	**.73**	**.71**	**.54**	−.03	−.07	−.05	−.11	.58	.62	.55	.52
PD	**.76**	**.94**	**.85**	**.82**	.18	−.08	.00	.07	−.08	−.07	−.01	−.03	.73	.72	.71	.73
OF	**.97**	**.71**	**.75**	**.80**	−.05	.14	.11	.08	.01	.05	.06	−.04	.88	.72	.75	.69
DF	-	**.95**	**.99**	**1.00**	-	−.17	−.23	**−.34**	-	.07	.09	.16	-	.77	.79	.78
ORF	**.81**	**.61**	**.69**	**.66**	−.14	.09	−.06	.16	.06	.13	.12	−.03	.58	.57	.52	.57
SP	**.72**	**.88**	**.66**	**.74**	−.08	−.03	.24	.16	.21	.04	.03	.06	.68	.77	.75	.77
SC	**.34**	.20	.20	.14	**.44**	**.45**	**.50**	**.51**	.01	.16	.05	−.07	.54	.52	.48	.33
EC	-	**.46**	.24	−.02	-	−.01	.27	**.52**	-	.17	.13	.00	-	.32	.32	.26
AWF	.12	−.01	−	-	−.02	.31	-	-	**.44**	.20	-	-	.26	.20	-	-
SWF	.12	−.05	.17	.11	.13	.29	.06	.09	**.35**	.29	.25	.08	.29	.23	.18	.05
MPS	.00	.05	−.08	−.11	**.60**	**.51**	**.72**	**.71**	.26	**.32**	.28	**.31**	.63	.63	.76	.73
NO	.12	.04	−.14	−.21	.25	.24	**.56**	**.65**	**.38**	**.62**	**.44**	**.37**	.43	.66	.67	.63
MFA	−.02	−.03	.07	.13	−.11	−.09	−.12	−.02	**.94**	**.88**	**.92**	**.86**	.75	.67	.80	.81
MFS	.00	−.02	.04	.11	.06	−.03	−.01	.04	**.82**	**.93**	**.90**	**.82**	.72	.81	.84	.80
MFM	-	.07	.04	−.05	-	−.10	.01	−.07	-	**.87**	**.86**	**.91**	-	.73	.80	.74
LC	−.15	−.01	−.03	.27	**.86**	**.79**	**.81**	**.61**	−.07	−.13	−.11	−.12	.53	.51	.52	.55
OE	−.06	−.08	.10	**.36**	**.85**	**.84**	**.72**	**.50**	.00	.00	−.05	−.08	.65	.62	.58	.54
PP	**.36**	**.57**	**.46**	**.55**	**.49**	.22	**.33**	.21	−.02	−.06	−.06	.09	.61	.50	.48	.56
OC	**.73**	**.73**	**.47**	**.54**	−.02	−.03	.24	.13	.09	.00	.05	.21	.60	.50	.49	.56

Note. Factor loadings greater than 1 suggest Heywood cases. [1]—Communalities represent the final estimates from the unrotated solution; WR—Word Reading; RC—Reading Comprehension; PD—Pseudoword Decoding; OF—Orthographic Fluency; DF—Decoding Fluency; ORF—Oral Reading Fluency; SP—Spelling; SC—Sentence Composition; EC—Essay Composition; AWF—Alphabet Writing Fluency; SWF—Sentence Writing Fluency; MPS—Math Problem Solving; NO—Numerical Operations; MFA—Math Fluency Addition; MFS—Math Fluency Subtraction; MFM—Math Fluency Multiplication; LC—Listening Comprehension; OE—Oral Expression; PP—Phonemic Proficiency; OC—Orthographic Choice; Columns represent age groups, 4 to 7, 8 to 11, 12 to 19, 20 to 50. **Bold** coefficients are greater than .30.

Table A7. Bifactor Solution with Three Specific Factors Across Age Groups.

	General				Factor 1				Factor 2				Factor 3				Communalities			
	4–7	**8–11**	**12–19**	**20–50**	**4–7**	**8–11**	**12–19**	**20–50**	**4–7**	**8–11**	**12–19**	**20–50**	**4–7**	**8–11**	**12–19**	**20–50**	**4–7**	**8–11**	**12–19**	**20–50**
WR	**−.92**	**.87**	**.85**	**.87**	.03	−.03	−.04	.18	−.06	.17	.13	−.13	.25	.17	.18	.06	.92	.82	.77	.81
RC	**−.72**	**.69**	**.66**	**.67**	−.26	**.36**	**−.33**	−.27	−.04	.01	.01	.00	.00	−.06	−.06	−.17	.58	.62	.55	.55
PD	**−.84**	**.79**	**.80**	**.85**	−.07	−.14	.16	.12	−.10	.12	.07	−.07	.06	.25	.19	.05	.73	.72	.71	.74
OF	**−.91**	**.83**	**.83**	**.82**	.06	.05	−.03	.02	−.03	−.03	−.02	−.06	.20	.29	**.32**	.11	.87	.77	.78	.69
DF	-	**.81**	**.78**	**.79**	-	−.15	.17	.26	-	−.01	−.01	.02	-	**.40**	**.42**	.29	-	.84	.82	.78
ORF	**−.73**	**.73**	**.67**	**.76**	.10	.03	.09	−.09	.02	−.09	−.05	−.03	.21	.27	.25	.08	.59	.62	.52	.58
SP	**−.84**	**.87**	**.87**	**.88**	.14	−.22	.09	.16	.04	.07	.05	.02	−.08	.01	.02	−.05	.73	.81	.77	.80
SC	**−.70**	**.73**	**.70**	**.49**	−.23	.07	−.03	−.11	−.05	−.07	.01	.02	−.14	**−.40**	−.26	−.25	.57	.70	.56	.32
EC	-	**.61**	**.59**	**.38**	-	**−.30**	.08	−.23	-	−.02	−.04	.10	-	**−.35**	−.20	−.24	-	.58	.40	.26
AWF	**−.44**	**.39**	-	-	.06	.11	-	-	.24	−.15	-	-	−.20	−.17	-	-	.29	.22	-	-
SWF	**−.53**	**.37**	**.37**	.22	.01	.18	−.07	−.30	.14	−.26	−.19	.12	**−.39**	−.01	.12	.13	.45	.24	.19	.17
MPS	**−.64**	**.69**	**.75**	**.58**	**−.44**	.26	**−.32**	−.04	.24	−.28	−.23	**.38**	.06	−.08	−.15	**−.55**	.66	.62	.75	.78
NO	**−.58**	**.65**	**.69**	**.46**	−.16	.07	−.18	.04	.26	**−.47**	**−.33**	**.42**	−.02	−.16	−.25	**−.60**	.43	.67	.68	.76
MFA	**−.58**	**.45**	**.59**	**.52**	.03	−.02	.04	−.02	**.63**	**−.69**	**−.67**	**.75**	−.05	.02	.05	.01	.74	.68	.80	.83
MFS	**−.63**	**.54**	**.64**	**.53**	−.09	.01	−.02	−.09	**.60**	**−.72**	**−.66**	**.75**	.02	.00	.03	.01	.76	.82	.84	.85
MFM	-	**.53**	**.65**	**.34**	-	-.05	.02	.11	-	**−.67**	**−.62**	**.78**	−	.02	−.04	−.08	-	.73	.80	.73
LC	**−.44**	**.56**	**.58**	**.67**	**−.60**	**.51**	**−.56**	**−.35**	.02	.02	.02	.01	.02	.04	.03	−.18	.56	.57	.64	.61
OE	**−.60**	**.62**	**.68**	**.70**	**−.53**	**.46**	**−.35**	−.19	.02	−.07	.01	.01	−.10	−.10	−.04	−.17	.65	.61	.59	.55
PP	**−.74**	**.70**	**.69**	**.74**	−.27	.01	−.04	−.13	-.05	.08	.07	.11	−.09	.01	.03	.04	.63	.50	.48	.58
OC	**−.75**	**.68**	**.71**	**.73**	.00	−.16	.05	.16	.07	.07	.01	.16	**.30**	.06	−.02	−.08	.65	.50	.51	.58
ω/ω.h	.90	.91	.92	.87	.03	.01	.01	.00	.03	.05	.03	.07	.00	.00	.00	.02				
H	.96	.96	.96	.96	.60	.53	.49	.39	.59	.77	.71	.82	.34	.46	.41	.58				
SS Loadings	8.23	9.00	9.29	8.23	1.10	.01	.76	.59	.99	.05	1.49	2.14	.48	.00	.63	1.02				
Prop Tot Variance	.48	.45	.49	.43	.06	.53	.04	.03	.06	.77	.08	.11	.03	.46	.03	.05				
Prop Comm Var	.76	.71	.76	.69	.10	.84	.06	.05	.09	.93	.12	.18	.04	.84	.05	.08				

Note. WR—Word Reading; RC—Reading Comprehension; PD—Pseudoword Decoding; OF—Orthographic Fluency; DF—Decoding Fluency; ORF—Oral Reading Fluency; SP—Spelling; SC—Sentence Composition; EC—Essay Composition; AWF—Alphabet Writing Fluency; SWF—Sentence Writing Fluency; MPS—Math Problem Solving; NO—Numerical Operations; MFA—Math Fluency Addition; MFS—Math Fluency Subtraction; MFM—Math Fluency Multiplication; LC—Listening Comprehension; OE—Oral Expression; PP—Phonemic Proficiency; OC—Orthographic Choice; Columns represent age groups, 4 to 7, 8 to 11, 12 to 19, 20 to 50. **Bold** coefficients are greater than .30.

Notes

[1] For convenience, we use the term *authors* throughout the article instead of the more accurate term *construction agency*. Nearly all modern standardized instruments are created by a team of people with different specialty knowledge (e.g., content matter, test construction techniques, item analysis), only a portion of which are credited on instrument documentation.

[2] We consider behavior to be a subclass of doings (Maraun 2013).

[3] Functional unity is applicable to phenomena from a variety of disciplines and knowledge domains, so may involve things other than behavior (e.g., neural activity).

[4] Technically, we *classify* attributes represented on a nominal scale rather than measure them. Classification has some properties similar to measurement, but they are distinct processes.

[5] The WIAT-4 authors provide some utility evidence for the Dyslexia Index score, but do not describe how they gathered this evidence in any detail (Breaux 2020, p. 114).

[6] Steps 2a and 2b are combined into a single Step 2 in the technical manual.

[7] Idiosyncratic employments of the intelligence concept continued throughout the 20th century and continue today (Legg and Hutter 2007).

[8] The bi-factor rotation requires extracting $p + 1$ factors, with the p indicating the number of group factors and *+1* indicating the additional general factor. Thus, we actually extracted 4–6 factors.

[9] We use the term *triadic theory* instead of the more common *gf –gc theory.* The latter term once had a specific meaning, but now it is more ambiguous as it can refer to either the theory Raymond Cattell created to extend Spearman's nöegenetic theory or the refinements and expansions to *gf –gc* theory initiated by Cattell's student, John Horn. Although Horn and Cattell worked together occasionally throughout Cattell's life, by the 1970s they had independent research programs and had developed separate intelligence theories. Thus, except for historical purposes, *gf –gc* theory is no longer viable because it has been replaced with two competing theories: Horn's *extended Gf–Gc theory* and Cattell's *triadic theory*.

[10] The acronym IDEA stands for the Individuals with Disabilities Education Act, which is an American law passed in 2004.

[11] The term *orthographic lexicon* is a more technical term for sight vocabulary (i.e., words we can correctly read instantly without effort).

[12] Promax rotation is oblique, meaning it allows the factors to be correlated.

[13] There are other ways for the classes to be ordered, but since we are employing common sense/intuitive meanings, we will not differentiate among them (for more details, see Michell 1999).

[14] Measurement models guaranteeing the conditions are not necessary, but the WIAT-4 authors do not discuss alternative probabilistic models in the technical manual.

[15] Some WIAT-4 subtests offer partial credit, so the raw scores would be the number of points earned.

[16] There are 495 ways to combine 4 out of 12 objects (i.e., ${}_{12}C_4$).

[17] The WIAT-4 provides both age- and grade-based norm groups for the norm-referenced scores, so it is likely more accurate to state the WIAT-4 provides 11 different score units in addition to the raw score.

[18] The WIAT-4 standard scores are all integers, so the values from Equation (1) must be rounded. The WIAT-4 authors do not provide information about the rounding function they employ, however, so we do not include one in Equation (1).

[19] The WIAT-4 authors hint, but do not state explicitly, that they normalized the raw score values within a norm group before converting to standard scores. It is true that normalizing can make a score's values have certain statistical properties, but it does not follow that the attribute the score values represent gains properties because of normalizing (Michell 2020; Thomas 1982). Thus, normalizing the raw scores does not change our evaluation that the WIAT-4 authors do not provide sufficient support for their claim that standard scores are on an equal-interval measurement scale.

References

Academic. 2021. Oxford English Dictionary Online. Available online: www.oed.com/view/Entry/880 (accessed on 10 October 2021).

Achievement. 2021. Oxford English Dictionary Online. Available online: www.oed.com/view/Entry/1482 (accessed on 10 October 2021).

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*, 4th ed. Washington, DC: American Educational Research Association.

Anastasi, Anne. 1976. *Psychological Testing*, 4th ed. New York: Macmillan.

Anastasi, Anne. 1980. Abilities and the measurement of achievement. *New Directions for Testing & Measurement* 5: 1–10.

Anastasi, Anne. 1984. Aptitude and achievement tests: The curious case of the indestructible strawperson. In *Social and Technical Issues in Testing: Implications for Test Construction and Use*. Edited by Barbara S. Plake. Hillsdale: Lawrence Erlbaum, Associates, pp. 129–40.

Bardos, Achilles N. 2020. *Basic Achievement Skills Inventory Comprehensive Test*, 2nd ed. Greeley: Edumetrisis.

Barrett, Paul T. 2018. The EFPA test-review model: When good intentions meet a methodological thought disorder. *Behavioral Sciences* 8: 5. [CrossRef]
Bascom, John. 1878. *Comparative Psychology: Or, the Growth and Grades of Intelligence*. New York: G. P. Putnam's Sons.
Beaujean, A. Alexander. 2015a. Adopting a new test edition: Psychometric and practical considerations. *Research and Practice in the Schools* 3: 51–57.
Beaujean, A. Alexander. 2015b. John Carroll's views on intelligence: Bi-factor vs. higher- order models. *Journal of Intelligence* 3: 121–36. [CrossRef]
Bem, Sacha, and Huib Looren De Jong. 2013. *Theoretical Issues in Psychology: An Introduction*, 3rd ed. Los Angeles: Sage.
Bennett, Maxwell R., and Peter Michael Stephan Hacker. 2022. *Philosophical Foundations of Neuroscience*, 2nd ed. Hoboken: John Wiley & Sons.
Berg, Irwin. A. 1959. The unimportance of test item content. In *Objective Approaches to Personality Assessment*. Edited by Bernard M. Bass and Irwin A. Berg. New York: Van Nostrand, pp. 83–99.
Berninger, Virginia W., and William D. Winn. 2006. Implications of advancements in brain research and technology for writing development, writing instruction, and educational evolution. In *Handbook of Writing Research*. Edited by Charles MacArthur, Steve Graham and Jill Fitzgerald. New York: Guilford Press, pp. 96–114.
Berninger, Virginia W., Donald. T. Mizokawa, Russell Bragg, Ana Cartwright, and Cheryl Yates. 1994. Intraindividual differences in levels of written language. *Reading and Writing Quarterly* 10: 259–75. [CrossRef]
Bollen, Kenneth A., and Shawn Bauldry. 2011. Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods* 16: 265–84. [CrossRef]
Boring, Edwin G. 1920. The logic of the normal law of error in mental measurement. *The American Journal of Psychology* 31: 1–33. [CrossRef]
Borsboom, Denny. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press.
Borsboom, Denny, Gideon J. Mellenbergh, and Jaap Van Heerden. 2004. The concept of validity. *Psychological Review* 111: 1061–71. [CrossRef]
Bradford, Gwen. 2016. Achievement, wellbeing, and value. *Philosophy Compass* 11: 795–803. [CrossRef]
Breaux, Kristina C. 2020. *Wechsler Individual Achievement Test: Technical & Interpretive Manual*, 4th ed. Bloomington: NCS Pearson.
Burisch, Matthias. 1984. Approaches to personality inventory construction: A comparison of merits. *American Psychologist* 39: 214–27. [CrossRef]
Buros, Oscar K. 1977. Fifty years in testing: Some reminiscences, criticisms, and suggestions. *Educational Researcher* 6: 9–15. [CrossRef]
Burt, Cyril L. 1917. *The Distributions and Relations of Educational Abilities*. London: Darling & Son.
Burt, Cyril L. 1944. Mental abilities and mental factors. *British Journal of Educational Psychology* 14: 85–94. [CrossRef]
Bush, Shane S., Jerry J. Sweet, Kevin J. Bianchini, Doug Johnson-Greene, Pamela M. Dean, and Mike R. Schoenberg. 2018. Deciding to adopt revised and new psychological and neuropsychological tests: An inter-organizational position paper. *The Clinical Neuropsychologist* 32: 319–25. [CrossRef]
Campbell, Jonathan M., Ronald T. Brown, Sarah E. Cavanagh, Ssarah F. Vess, and Mathew J. Segall. 2008. Evidence-based assessment of cognitive functioning in pediatric psychology. *Journal of Pediatric Psychology* 33: 999–1014. [CrossRef]
Carroll, John B. 1943. The factorial representation of mental ability and academic achievement. *Educational and Psychological Measurement* 3: 307–31. [CrossRef]
Carroll, John B. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge: Cambridge University Press.
Cattell, Raymond B. 1944. Psychological measurement: Normative, ipsative, interactive. *Psychological Review* 51: 292–303. [CrossRef]
Cattell, Raymond B. 1956. Personality and motivation theory based on structural measurement. In *Psychology of Personality: Six Modern Approaches*. Edited by James L. McCary. New York: Logos Press, pp. 63–119.
Cattell, Raymond B. 1987. *Intelligence: Its Structure, Growth, and Action*. Amsterdam: Elsevier.
Cattell, Raymond B., and Ronald C. Johnson, eds. 1986. *Functional Psychological Testing: Principles and Instruments*. New York: Brunner/Maze.
Coffman, William E. 1970. Concepts of achievement and proficiency. In *1969 Invitational Conference on Testing Problems: Toward a Theory of Achievement Measurement*. Edited by Philip H. DuBois. Princeton: Educational Testing Service, pp. 3–11.
Coombs, Clyde H. 1948. Some hypotheses for the analysis of qualitative variables. *Psychological Review* 55: 167–74. [CrossRef]
Coombs, Clyde. H., Howard Raiffa, and Robert M. Thrall. 1954. Mathematical models and measurement theory. In *Decision Processes*. Edited by Robert M. Thrall, Clyde Hamilton Coombs and Robert L. Davis. New York: Wiley, pp. 19–37.
Crandall, Vaughn J. 1963. Achievement. In *The Sixty-Second Yearbook of the National Society for the Study of Education, Part 1: Child Psychology*. Edited by H. W. Stevenson, J. Kagan and C. Spiker. Chicago: University of Chicago Press, pp. 416–59.
Cronbach, Lee J. 1990. *Essentials of Psychological Testing*, 5th ed. London: Harper Collins.
Dailey, John Thomas, and Marion F. Shaycoft. 1961. *Types of Tests in Project Talent: Standardized Aptitude and Achievement Tests*. Cooperative Research Monograph No. 9. Washington, DC: United States Government Printing Office.
Danziger, Kurt. 1997. *Naming the Mind: How Psychology Found Its Language*. Thousand Oaks: Sage.
Ebel, Robert L., and David A. Frisbie. 1991. *Essentials of Educational Measures*, 5th ed. New Deli: Prentice-Hall of India.
Edwards, Jeffrey R. 2011. The fallacy of formative measurement. *Organizational Research Methods* 14: 370–88. [CrossRef]

Evers, Arne, Carmen Hagemeister, Andreas Høstmælingen, Patricia A. Lindley, José Muñiz, and Anders Sjöberg. 2013. *EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests*. Version 4.2.6. Brussels: European Federation of Psychology Associations. Available online: http://assessment.efpa.eu/documents-/ (accessed on 10 October 2021).
French, John W. 1951. *The Description of Aptitude and Achievement Tests in Terms of Rotated Factors*. Chicago: University of Chicago Press.
Goodey, Christopher F. 2011. *A History of Intelligence and "Intellectual Disability": The Shaping of Psychology in Early Modern Europe*. Farnham: Ashgate Publishing.
Guilford, Joy Paul. 1946. New standards for test evaluation. *Educational and Psychological Measurement* 6: 427–38. [CrossRef]
Guttman, Louis. 1977. What is not what in statistics. *Journal of the Royal Statistical Society. Series D (The Statistician)* 26: 81–107. [CrossRef]
Guttman, Louis, and Shlomit Levy. 1991. Two structural laws for intelligence tests. *Intelligence* 15: 79–103. [CrossRef]
Hacker, Peter Michael Stephan. 2013. *The Intellectual Powers: A Study of Human Nature*. Malden: Wiley-Blackwell.
Hacker, Peter, and Michael Stephan. 2020. Methods of connective analysis. In *Philosophy in the Age of Science?: Inquiries into Philosophical Progress, Method, and Societal Relevance*. Edited by Julia Hermann, Jeroen Hopster, Wouter Kalf and Michael Klenk. London: Rowman & Littlefield, pp. 111–30.
Hand, David J. 2004. *Measurement Theory and Practice: The World through Quantification*. London: Edward Arnold.
Hand, David J. 2016. *Measurement: A Very Short Introduction*. Oxford: Oxford University Press.
Hardy, Ben. 2009. Morale: Definitions, Dimensions and Measurement. Ph.D. dissertation, University of Cambridge, Cambridge, UK. Available online: https://www.repository.cam.ac.uk/handle/1810/229514 (accessed on 10 October 2021).
Haynes, Stephen N., David C. S. Richard, and Edward S. Kubany. 1995. Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment* 7: 238–47. [CrossRef]
Hearnshaw, Leslie Spencer. 1941. Psychology and operationism. *Australasian Journal of Psychology and Philosophy* 19: 44–57. [CrossRef]
Heckhausen, Heinz. 1967. *The Anatomy of Achievement Motivation*. Translated by Kay F. Butler, Robert C. Birney, and David C. Mcclelland. Cambridge: Academic Press.
Holzinger, Karl John, Frances Swineford, and Harry H. Harman. 1937. *Student Manual of Factor Analysis: An Elementary Exposition of the Bi-Factor Method and Its Relation to Multiple-Factor Methods*. Chicago: University of Chicago Department of Education.
Hoover, Wesley A., and Philip B. Gough. 1990. The simple view of reading. *Reading and Writing* 2: 127–60. [CrossRef]
Horn, John L. 1963. The discovery of personality traits. *The Journal of Educational Research* 56: 460–65. [CrossRef]
Horn, John L. 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika* 30: 179–85. [CrossRef]
Horn, John L. 1972. State, trait and change dimensions of intelligence. *British Journal of Educational Psychology* 42: 159–85. [CrossRef]
Howard, Robert W. 1993. On what intelligence is. *British Journal of Psychology* 84: 27–37. [CrossRef]
International Test Commission. 2001. International guidelines for test use. *International Journal of Testing* 1: 93–114. [CrossRef]
Jennrich, Robert I., and Peter M. Bentler. 2011. Exploratory bi-factor analysis. *Psychometrika* 76: 537–49. [CrossRef]
Johnson, James M., Henry S. Pennypacker, and Gina Green. 2019. *Strategies and Tactics of Behavioral Research and Practice*, 4th ed. New York: Routledge.
Joint Committee for Guides in Metrology. 2012. JCGM 200:2012. International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM). Available online: https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf (accessed on 22 October 2021).
Kaiser, Henry F. 1974. An index of factorial simplicity. *Psychometrika* 39: 31–36. [CrossRef]
Kane, Michael T. 2013. Validating the interpretations and uses of test scores. *Journal of Educational Measurement* 50: 1–73. [CrossRef]
Kaufman, Scott Barry, Matthew R. Reynolds, Xin Liu, Alan S. Kaufman, and Kevin S. McGrew. 2012. Are cognitive *g* and academic achievement *g* one and the same *g*? An exploration on the Woodcock-Johnson and Kaufman tests. *Intelligence* 40: 123–38. [CrossRef]
Kaufman, Allen S., Nadeen L. Kaufman, and Kristina C. Breaux. 2014. *Kaufman Test of Educational Achievement (3rd ed) Technical & Interpretive Manual*. Bloomington: NCS Pearson.
Kilpatrick, David A. 2015. *Essentials of Assessing, Preventing, and Overcoming Reading Difficulties*. Hoboken: Wiley.
Kim, Young-Suk, Brandy Gatlin, Stephanie Al Otaiba, and Jeanne Wanzek. 2018. Theorization and an empirical investigation of the component-based and developmental text writing fluency construct. *Journal of Learning Disabilities* 51: 320–35. [CrossRef]
Kline, Paul. 1998. *The New Psychometrics: Science, Psychology, and Measurement*. London: Routledge.
Krause, Merton S. 1967. The construct validity of measuring instruments. *The Journal of General Psychology* 77: 277–84. [CrossRef]
Krause, Merton S. 2005. How the psychotherapy research community must work toward measurement validity and why. *Journal of Clinical Psychology* 61: 269–83. [CrossRef]
Krause, Merton S. 2012. Measurement validity is fundamentally a matter of definition, not correlation. *Review of General Psychology* 16: 391–400. [CrossRef]
Legg, Shane, and Marcus Hutter. 2007. A collection of definitions of intelligence. In *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*. Edited by Ben Goertzel and Pei Wang. Amsterdam: IOS Press, pp. 17–24.
Lindquist, E. F. 1936. The theory of test construction. In *The Construction and Use of Achievement Examinations: A Manualfor Secondary School Teachers*. Edited by Herbert E. Hawkes, E. F. Lindquist and C. R. Mann. Boston: Houghton Mifflin, pp. 17–106.
Loehlin, John C., and A. Alexander Beaujean. 2016a. *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis*, 5th ed. London: Routledge.

Loehlin, John C., and A. Alexander Beaujean. 2016b. *Syntax Companion for Latent Variable Models: An Introduction to Factor, Path, And Structural Equation Analysis*, 5th ed. Waco: Baylor Psychometric Laboratory.

Loevinger, Jane. 1947. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs* 61: i-49. [CrossRef]

Loevinger, Jane. 1957. Objective tests as instruments of psychological theory. *Psychological Reports* 3: 635–94. [CrossRef]

Lykken, David T. 1968. Statistical significance in psychological research. *Psychological Bulletin* 70: 151–59. [CrossRef]

Maraun, Michael D. 1998. Measurement as a normative practice: Implications of Wittgenstein's philosophy for measurement in psychology. *Theory & Psychology* 8: 435–61. [CrossRef]

Maraun, Michael D. 2013. The concepts of suicidology. In *A Wittgensteinian Perspective on the Use of Conceptual Analysis in Psychology*. Edited by Timothy P. Racine and Kathleen L. Slaney. London: Palgrave Macmillan, pp. 233–52.

Mari, Luca, Paolo Carbone, and Dario Petri. 2015. Fundamentals of hard and soft measurement. In *Modern Measurements: Fundamentals and Applications*. Edited by Alessandro Ferrero, Dario Petri, Paolo Carbone and Marcantonio Catelani. Hoboken: Wiley-IEEE Press, pp. 203–62.

Mari, Luca, Andrew Maul, David Torres Irribarra, and Mark Wilson. 2017. Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement* 100: 115–21. [CrossRef]

Markus, Keith A., and Denny Borsboom. 2013. *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. New York: Routledge.

Mather, Nancy, and Bashir Abu-Hamour. 2013. Individual assessment of academic achievement. In *APA Handbook of Testing and Assessment in Psychology, Vol. 3: Testing and Assessment in School Psychology and Education*. Edited by Kurt F. Geisinger, Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise and Michael C. Rodriguez. Washington, DC: American Psychological Association, pp. 101–28.

Matsumoto, David, ed. 2009. *The Cambridge Dictionary of Psychology*. Cambridge: Cambridge University Press.

McFall, Richard M., and Teresa A. Treat. 1999. Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology* 50: 215–41. [CrossRef]

Merton, Robert K. 1968. *Social Theory and Social Structure*, Enlarged ed. New York: Free Press.

Messick, Samuel. 1989. Validity. In *Educational Measurement*, 3rd ed. Edited by Robert Linn. Washington, DC: American Council on Education, pp. 13–103.

Michell, Joel. 1990. *An Introduction to the Logic of Psychological Measurement*. Hillsdale: Erlbaum.

Michell, Joel. 1999. *Measurement in Psychology: Critical History of a Methodological Concept*. Cambridge: Cambridge University Press.

Michell, Joel. 2005. The logic of measurement: A realist overview. *Measurement* 38: 285–94. [CrossRef]

Michell, Joel. 2009. Invalidity in validity. In *The Concept of Validity: Revisions, New Directions, and Applications*. Edited by Robert W. Lissitz. Charlotte: IAP Information Age Publishing, pp. 111–33.

Michell, Joel. 2020. Thorndike's credo: Metaphysics in psychometrics. *Theory & Psychology* 30: 309–28. [CrossRef]

Mitchell, James V., Jr. 1984. Testing and the Oscar Buros lament: From knowledge to implementation to use. In *Social and Technical Issues in Testing: Implications for Test Construction and Usage*. Edited by Barbara S. Plake. Hillsdale: Erlbaum, pp. 111–26.

Monroe, Walter S., James C. DeVoss, and George W. Reagan. 1930. *Educational Psychology*. Garden City, NY: Doubleday, Doran & Company.

NCS Pearson. 2020. *Wechsler Individual Achievement Test*, 4th ed. Bloomington: Author.

Newton, Paul E. 2017. There is more to educational measurement than measuring: The importance of embracing purpose pluralism. *Educational Measurement: Issues and Practice* 36: 5–15. [CrossRef]

Newton, Paul E., and Stuart D. Shaw. 2014. *Validity in Educational and Psychological Assessment*. Thousand Oaks, CA: Sage.

Norenzayan, Ara, and Steven J. Heine. 2005. Psychological universals: What are they and how can we know? *Psychological Bulletin* 131: 763–84. [CrossRef]

Ozer, Daniel J., and Steven P. Reise. 1994. Personality assessment. *Annual Review of Psychology* 45: 357–88. [CrossRef]

Peak, H. 1953. Problems of objective observation. In *Research Methods in the Behavioral Sciences*. Edited by Leon Festinger and Daniel Katz. New York: Dryden Press, pp. 243–99.

R Development Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reeve, Charlie L., and Silvia Bonaccio. 2011. The nature and structure of "intelligence". In *The Wiley-Blackwell Handbook of Individual Differences*. Edited by Tomas Chamorro-Premuzic, Sophie von Stumm and Adrian Furnham. Hoboken: Wiley Blackwell, pp. 187–216.

Reynolds, Cecil. R. 1998. Fundamentals of measurement and assessment in psychology. In *Comprehensive Clinical Psychology: Vol. 4: Assessment*. Edited by Cecil R. Reynolds. New York: Pergamon/Elsevier, pp. 33–55.

Rhemtulla, Mijke, Riet van Bork, and Denny Borsboom. 2015. Calling models with causal indicators "measurement models" implies more than they can deliver. *Measurement: Interdisciplinary Research and Perspectives* 13: 59–62. [CrossRef]

Rugg, H. O., ed. 1921. Intelligence and its measurement: A symposium. [Special issue]. *Journal of Educational Psychology* 12: 123–47. [CrossRef]

Schneider, W. Joel. 2013. Principles of assessment of aptitude and achievement. In *The Oxford Handbook of Child Psychological Assessment*. Edited by Donald H. Saklofske, Cecil R. Reynolds and Vicki L. Schwean. New York: Oxford University Press, pp. 286–330. [CrossRef]
Shapiro, Edward S. 2011. *Academic Skills Problems: Direct Assessment and Intervention*, 4th ed. New York: Guilford.
Sijtsma, Klaas. 2006. Psychometrics in psychological research: Role model or partner in science? *Psychometrika* 71: 451. [CrossRef] [PubMed]
Slaney, Kathleen. 2017. *Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions*. London: Palgrave.
Spearman, Charles E. 1927. *The Abilities of Man: Their Nature and Measurement*. London: Macmillan.
Spearman, Charles E. 1933. The factor theory and its troubles. III. Misrepresentation of the theory. *Journal of Educational Psychology* 24: 591–601. [CrossRef]
Spearman, Charles E. 1937. *Psychology Down the Ages*. Oxford: Macmillan, vol. 1.
Spearman, Charles E. 1938. Measurement of intelligence. *Scientia, Milano* 64: 75–82. [CrossRef]
Spearman, Carlesta Elliot, and Llewellyn Wynn Jones. 1950. *Human Ability: A Continuation of "The Abilities of Man"*. London: Macmillan.
Spinath, Birgit. 2012. Academic achievement. In *Encyclopedia of Human Behavior*, 2nd ed. Edited by Vilayanur S. Ramachandran. Cambridge: Academic Press, pp. 1–8.
Steiner, Markus D., and Silvia Grieder. 2020. EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software* 5: 2521. [CrossRef]
Stokes, Dustin. 2008. A metaphysics of creativity. In *New Waves in Aesthetics*. Edited by Kathleen Stock and Katherine Thomson-Jones. London: Palgrave Macmillan, pp. 105–24.
Taine, Hippolyte. 1872. *On Intelligence*, Rev ed. Translated by T. D. Haye. New York: Holt & Williams.
Thomas, Lawrence G. 1942. Mental tests as instruments of science. *Psychological Monographs* 54: i-87. [CrossRef]
Thomas, Hoben. 1982. IQ, interval scales, and normal distributions. *Psychological Bulletin* 91: 198–202. [CrossRef]
Thorndike, Robert M., and Tracy Thorndike-Christ. 2010. *Measurement and Evaluation in Psychology and Education*, 8th ed. London: Pearson.
Varzi, Achille C., and Giuliano Torrengo. 2006. Crimes and punishments. *Philosophia* 34: 395–404. [CrossRef]
Velicer, Wayne F. 1976. Determining the number of components from the matrix of partial correlations. *Psychometrika* 41: 321–27. [CrossRef]
Vendler, Zeno. 1957. Verbs and times. *The Philosophical Review* 66: 143–60. [CrossRef]
Vernon, Philip E. 1950. *The Structure of Human Abilities*. New York: Wiley.
Vygotsky, Lev. S. 1987. The historical meaning of the crisis in psychology: A methodological investigation. In *The Collected Works of L. S. Vygotsky, Vol. 3: Problems of the Theory and History of Psychology*. Edited by Robert W. Rieber and Jeffre Wollock. Translated by René van Der Veer. New York: Springer, pp. 233–343.
Wechsler, David. 1975. Intelligence defined and undefined: A relativistic appraisal. *American Psychologist* 30: 135–39. [CrossRef]
Wechsler, David, Susan E. Raiford, and James A. Holdnack. 2014. *Wechsler Intelligence Scale for Children-Fifth Edition: Technical and Interpretive Manual*. Bloomington, MN: NCS Pearson.
Wesman, Alexander G. 1956. *Aptitude, Intelligence, and Achievement*. Test Service Bulletin 51. New York: The Psychological Corporation.
Woodcock, Richard W. 1999. What can Rasch-based scores convey about a person's test performance. In *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. Edited by Susan E. Embretson and Scott L. Hershberger. Hoboken: Erlbaum, pp. 105–27.
Zimprich, Daniel, and Mike Martin. 2009. A multilevel factor analysis perspective on intellectual development in old age. In *Aging and Cognition: Research Methodologies and Empirical Advances*. Edited by Hayden B. Bosworth and Christopher Hertzog. Washington, DC: American Psychological Association, pp. 53–76.

Journal of
Intelligence

MDPI

Article

The Impact of an Enrichment Program on the Emirati Verbally Gifted Children

Hala Elhoweris [1], Najwa Alhosani [2], Negmeldin Alsheikh [2], Rhoda-Myra Garces Bacsal [1] and Eleni Bonti [1,3,4,*]

[1] Special Education Department, College of Education, United Arab Emirates University, Al Ain 112612, United Arab Emirates
[2] College of Education, United Arab Emirates University, Al Ain 112612, United Arab Emirates
[3] First Psychiatric Clinic, School of Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki, 'Papageorgiou' General Hospital of Thessaloniki, Agiou Pavlou 76, Pavlos Melas, 56429 Thessaloniki, Greece
[4] Department of Education, School of Education, University of Nicosia, Nicosia 2417, Cyprus
* Correspondence: bonti@auth.gr

Abstract: Most researchers agree that verbally gifted learners should be provided with differentiated curriculum experiences that will allow them to reach their full potential. However, research is scarce in the field. The present study examined the impact of a reading enrichment program on fourth-grade students' critical reading abilities. The program was based on the Integrated Curriculum Model (ICM). The sample consisted of forty fourth-grade verbally gifted students from a school in Dubai, who were randomly assigned to either an experimental instruction condition or a traditional instruction condition and completed pre and post-tests of language arts. A pre-and post-experimental design was used. The overall results indicated the efficacy of the differentiated enrichment program in enhancing Emirati gifted learners' critical reading abilities. The study also provides a framework for better provision and teacher training planning regarding gifted education in the UAE.

Keywords: verbally gifted learners; Integrated Curriculum Model (ICM); UAE; reading enrichment programs; language arts

Citation: Elhoweris, Hala, Najwa Alhosani, Negmeldin Alsheikh, Rhoda-Myra Garces Bacsal, and Eleni Bonti. 2022. The Impact of an Enrichment Program on the Emirati Verbally Gifted Children. *Journal of Intelligence* 10: 68. https://doi.org/10.3390/jintelligence10030068

Received: 7 May 2022
Accepted: 29 June 2022
Published: 15 September 2022

"Giftedness is arguably the most precious natural resource a civilization can have"
(Sternberg and Davidson, as cited in Pfeiffer 2002, p. 32).

1. Introduction: The Impact of an Enrichment Program on the Emirati Verbally Gifted Children

1.1. The Concept of Giftedness

Giftedness has traditionally been conceptualized based on high performance on IQ tests. In that sense, it is often related to a commonly shared underlying, one-dimensional conception of a high intelligence quotient (IQ), often referred to as Spearman's 'g' factor, which is measured by psychometric instruments (Beckmann and Minnaert 2018). The term 'high achievement' has also been used as a synonym for giftedness, both of which describe "students who consistently produce ideas and/or products of excellence" (Jiboye et al. 2019).

The federal Elementary and Secondary Education Act defines giftedness as: "students, children, or youth who give evidence of high achievement capability in areas such as intellectual, creative, artistic, or leadership capacity, or in specific academic fields, and who need services and activities not ordinarily provided by the school, in order to fully develop those capabilities" [ESEA, (Paul 2015)]. Several multi-dimensional models of giftedness have been developed through the years, which include various other domains or aspects of intelligence (Renzulli 1978; Monks and Mason 2000).

Furthermore, there is evidence of insufficient awareness of the definition of giftedness among teachers and parents. For instance, AlGhawi (2017) documented that, although there is an official definition of giftedness adopted by the Ministry of Education (MoE)

and published for schools in Dubai, teachers and parents in the UAE continue to define giftedness partially and differently. These findings offer further testimony to the general limited knowledge around the issues of both defining and identifying giftedness, let alone applying appropriate strategic plans for gifted education.

In summary, gifted children are considered to be a heterogeneous group of students, who manifest a high ability and/or talent in several domains (e.g., cognitive/mental, linguistic, etc.), along with multiple interpersonal characteristics (Monks and Mason 2000), and who require a differentiated curriculum to meet their unique abilities and needs.

1.2. Verbally Gifted Students—Critical and Creative Reading

The term 'verbally gifted' is used to refer to children who have significantly stronger language skills than their peers (Winnebrenner 2004). These skills allow the learner to achieve a full understanding of the text being read. VanTassel-Baska (2003) defines verbally gifted as "gifted children who achieve language competency at an earlier age than their chronological age-mates" (p. 1). Verbally gifted children typically present several characteristics, which are different from their peers, related to high verbal ability, early reading, advanced vocabulary, and high-level reading comprehension (Colangelo and Davis 2003).

1.3. The Need for Curriculum Differentiation and Educational Modifications for the Verbally Gifted

Gifted children have different educational needs than their counterparts, which cannot be met through curricula designed for their non-gifted peers (Scruggs and Cohn 1983; Wynn 1990; Maker 2005; Wood 2008). It is critical, therefore, to identify these children and provide them with the appropriate educational programs to meet their individual needs.

Dooley (1993) cautions that a stimulating reading program for gifted readers must have at least two major components. These include provision for a quick mastering of the basic curriculum through curriculum compacting, along with a differentiated curriculum, which should involve modifications of the content and the processes used to explore that content.

Finally, Davis et al. (2011) have stressed that a lack of recognition of gifted learners' educational and developmental needs, and a lack of appropriate accommodation provided, might put these children 'at risk' of failing to fully develop and flourish educationally.

2. Enrichment Models for Teaching Gifted Learners

The important 'ingredients' of instructional approaches that aid the cognitive development of gifted/talented (g/t) or high-ability students and the necessary alterations of the curriculum that should be made to enhance g/t students' learning experiences, have been broadly recognized and extensively discussed in the literature (Brown and Campione 1994; Csikszentmihalyi et al. 1993; McLaughlin and Talbert 1993; Newstead and Wason 1995; Vye et al. 1998).

However, in many schools around the world, gifted children are given the same quantity and quality academic work as their peers. As Renzulli (2005) points out, this could be a significant waste of their school time since gifted students need to be grouped with their gifted peers for enrichment activities. Indeed, as stated by Davis et al. (2011), "grouping for enrichment, either within the class or in a resource room (pullout program), produces substantial gains in academic achievement, creativity, and other thinking skills" (p. 13).

Various enrichment models and programs have been developed and widely implemented all over the world, to facilitate and reinforce gifted students' academic, social, creative, and thinking skills, and abilities and needs across several domains. Most of these models share an enriched view of curriculum development for the gifted, which addresses a broader conception of giftedness, taking into account principles of creativity, motivation, and independence as crucial constructs to the development of high ability. In addition, advanced process skills, such as critical thinking and creative problem solving are also viewed as central within these models (VanTassel-Baska and Brown 2007).

Some of the most well-known enrichment models include The Renzulli Schoolwide Enrichment Triad Model (SEM, Renzulli 1977; Renzulli et al. 1981) and Baska's Integrated Curriculum Model (VanTassel-Baska 1995).

2.1. The Integrated Curriculum Model (ICM)

Even though most of the aforementioned enrichment models aimed at meeting gifted learners' needs across various curriculum domains, some of these models specifically focused on promoting verbally gifted learners' abilities, in particular (e.g., SEM, Renzulli 2005; Maker Matrix, Maker 2005).

VanTassel-Baska's (1995) Integrated Curriculum Model (ICM) is one of the most extensively researched curriculum development models in gifted education, which also prioritizes verbal giftedness. ICM is based on differentiated instruction and includes forty units in science, language arts, social studies, and mathematics. The process of instruction and learning included in the ICM curriculum for verbally gifted students is modified in a variety of ways, including giving emphasis on higher levels of thinking and increasing the level of abstractness (Davis et al. 2011). In addition, the salient features of this curriculum include accelerated and advanced content, depth, and complexity through abstract concepts, direct study of higher-order thinking processes, interdisciplinary themes, and student research (Avery and Little 2003).

According to its developer, the ICM demonstrates the power of using a clear design approach. More precisely, its formula for a 'successful curriculum' is based on the fact that it couples linked subject-based standards with strong elements of differentiation for gifted learners (VanTassel-Baska and Little 2011).

VanTassel-Baska first proposed the Integrated Curriculum Model (ICM) in 1986, based on what worked with gifted learners, according to the relevant literature available at that time.

The theoretical origins of the ICM are grounded on the early conceptualizations of Vygotsky (1978), particularly on his notion of the zone of proximal development (ZPD). Other sources central to the ICM theoretical background include Csikszentmihalyi's (1991) concept of 'flow', according to which gifted learners demonstrate a broader and deeper capacity to engage in learning than their typical counterparts (Csikszentmihalyi et al. 1993), and the view of interactionism, whereby the learner increases learning depth by interacting with others to enhance understanding of concepts and ideas. The theory of constructivism whereby learners construct knowledge for themselves (i.e., they are in charge of their own learning), is also central to the instructional processes applied within the ICM curriculum.

The model was further expounded upon in the subsequent years (VanTassel-Baska 2008, 2015), and today it is comprised of three interrelated dimensions that are responsive to different aspects of the gifted learner: First, it emphasizes advanced content knowledge that frames disciplines of study through the use of advanced materials in each subject area and by altering the scope and sequence of curriculum to meet the needs of the gifted (VanTassel-Baska et al. 2000). Second, it provides higher-order thinking and processing and third, it organizes learning experiences around the important aspects of a discipline, leading to an in-depth understanding of each discipline, while also providing connections across disciplines. Taken together, these relatively distinct curriculum dimensions formed the basis of the ICM, and have proven successful with gifted learners at various developmental stages and across several domain-specific areas (VanTassel-Baska and Little 2011; VanTassel-Baska and Stambaugh 2006).

2.2. International Research on the Application of the ICM Model

Findings from several intervention studies in different countries around the world, which have used the ICM in teaching language arts to gifted learners, provide sufficient evidence that the particular enrichment model was successful in promoting verbally gifted students' knowledge, attitudes, motivation, and thinking skills (e.g., Brown et al. 2006; Feng et al. 2004; Gubbins et al. 2002; Kim et al. 2012; VanTassel-Baska and Brown 2007). Most of

these studies used quasi-experimental research designs, which compared the pre-test/post-test performance of gifted students participating in these programs, as well as with the performance of their gifted peers, who were taught using the mainstream curriculum. Based on the results of these studies, the ICM model has demonstrated its ability and effectiveness in providing an enhanced learning experience that allows optimal learning and development of gifted learners, especially in the areas of language arts (VanTassel-Baska et al. 2009). Additional research studies suggest that the language art enrichment programs for the verbally gifted learners should include an appropriate selection of reading materials, guided critical discussions and advanced organizers for processing, the use of broad themes and concepts, independent research, and interdisciplinary connections (Winnebrenner 2004). It should be noted, however, that only a few studies have been conducted that focused only on reading instructional programs for g/t students (Wood 2008). In an earlier study, positive changes in teachers' attitudes, student motivational response, and school district changes were documented as a result of implementing the ICM science and language arts curricula over three years (VanTassel-Baska et al. 2000). Another study conducted by Feng et al. (2004), examined the effectiveness of the ICM implementation in a suburban school district of g/t students in grades 3 to 5. Their results revealed significant levels of enhancement in gifted learners, in terms of language arts, critical reading, persuasive writing, and scientific research design skills.

2.3. Verbally Gifted Student's Education in the UAE

Gifted education has gained much popularity lately, as it has become a prominent issue in the Arabian Gulf. In response to the international calls for inclusive education as a form of equity in education, and following Merry's (2008) suggestion that gifted students worldwide should have justice in education, the UAE educational system acknowledged that gifted children have a right to be recognized and catered to within school. Hence, the 'School for All' initiative, in line with the MoE's Strategy 2010–2020, focused on encompassing all special needs services for both gifted students and students with disabilities. As a result, gifted education has been gaining momentum, interest, and support from ministries of education in these countries and government funds have been increasing. Additionally, in 2008, the MoE created the 'development of gifted and talented students' skills' initiative, which was joined by many schools for gifted learners in the UAE. Hundreds of gifted students benefited from this initiative, whilst many teachers received training on identification and intervention programs for the g/t students. In 2014, the MoE introduced a new initiative called the 'integrated system to identify and care for talents' (AlGhawi 2017). Since then, several other organizations implemented various programs for g/t students in the Arabic Emirates (e.g., the Hamdan Bin Rashid AlMaktoum Foundation for distinguished Performance, 2015; the Emirates Association for the Gifted; etc.) and various agencies were created to support gifted education in the Gulf (e.g., the 'Abu Dhabi Education Council' (Abu Dhabi Education Council 2011) and the 'Human Development Authority in Dubai' (Knowledge and Human Development Authority 2011).

In its continuous and keen efforts to excel academic performance in the country, the Ministry of Education (MoE) in the UAE launched an ambitious, strategic four-year (2017–2021), developmental plan that is wholly characterized by innovation and creativity. This aspiring education system aimed to instill a sought-after knowledge base that would produce competitiveness among society members speared at creating a distinguished realm at all venues.

The ultimate goal was to provide the UAE with the best possible human resource base, which would meet, and surpass, future market demands in various conventional and newly introduced contexts.

In addition, literacy has been and remains a cornerstone for the educational, social, economic, and personal fulfilment of UAE citizens. Indeed, according to the UAE Vision (2010), literate citizens in the UAE must be able to respond thoughtfully and articulately in oral and written forms, to fully participate in economic, political, social, and educational dialogues.

The emphasis recently given to literacy in the UAE has led educators and parents to question students' reading and writing achievements in the English and Arabic languages. Moreover, since 2008, the United Arab Emirates have participated in several international standardized tests to examine and benchmark the performance levels of its education system. These tests include the 'Program for International Student Assessment, (PISA) (OECD 2019, 2021), 'Trends in International Mathematics and Science Study (TIMSS), and the 'Progress in International Reading Literacy Study' (PIRLS) (Martin et al. 2007).

The UAE received top ranks in the Arab world, but unfortunately, results from the PISA released in 2016, showed that UAE students continue to fall below the 'Organization for Economic Co-operation and Development' (OECD) average in science, reading, and mathematics (OECD 2019, 2021). This undesired result was in great conflict with the UAE's National Agenda calls, according to which, the UAE was supposed to rank among the top 20 in PISA by 2021.

Meanwhile, there has been a virtual national 'panic' about reading and writing achievements in the UAE elementary, secondary, and postsecondary education. Indeed, there is a consensus that literacy levels nationally are unsatisfactory in the UAE (Ghefli 2016). This conceptualization, in combination with the discouraging results from PISA (OECD 2019), might have, however, functioned as a driving force behind reforming and improving language skills curricula with an increased focus on reading in the UAE. In addition, since reading is one of the most important skills students need to master at all academic levels, giving this topic the urgency and the absolute importance was very critical to the UAE education sector (UAE Innovation Strategy 2015).

Moreover, recent trends in critical literacy around the world have focused on critical reading and critical thinking, and the United Arab Emirates are no exception. Critical reading and critical thinking are the highest processes in reading and thinking, which entail the ability of "using careful evaluation, sound judgment, and reasoning powers" (Milan 1995, p. 218). Accordingly, the UAE emphasized the use of critical thinking and critical reading as a panacea for low language performance, as indicated by many local and international standardized tests; the impetus for that was to prepare students who can think critically, reason logically, evaluate different sources of information, and efficiently apply the learned knowledge in realistic conditions. Finally, the renewed and sustained economic growth in the United Arab Emirates and the overall well-being of all citizens in the Gulf countries led stakeholders to invest in high-quality learning (Elhoweris 2014).

Nevertheless, as AlGhawi (2017) argued, although gifted education has been adequately established in the USA and Europe (Davis and Rimm 2004), it is a relatively new initiative in the UAE. Therefore, there is a paucity of programs that address the unique educational needs of verbally gifted students. More precisely, to date, in the UAE public schools, no gifted and talented programs have been implemented that address the unique educational needs of verbally gifted students, to become productive and contributing members of society. In fact, verbally gifted students have been the most neglected group in the UAE public schools. Some of these students have not been even recognized as being gifted, because they do not write well or do not excel in all areas of language arts (AlGhawi 2017). This could be a tragic waste for them and the UAE society as well. As Davis et al. (2011) state, without appropriate education, gifted children could suffer psychological damage and permanent impairment of their abilities. Other researchers have also stressed the positive impact of challenging instruction on the emotional, affective, and social development of high-ability students (Eddles-Hirsch et al. 2010; Cross 2011).

In her study, which investigated the provision of gifted education in Dubai-UAE, using the National Association for Gifted Children (National Association for Gifted Children 2010) program standards and the implementation of gifted education programs in seven primary government schools in Dubai, AlGhawi (2017) revealed several shortcomings and discrepancies. More specifically, AlGhawi acknowledges that there has been a positive progression in gifted education during the 21st century; however, the findings of her recent study raised questions about the modes of implementation of gifted education in the UAE.

At the same time, the results of this study highlighted additional deficiencies concerning the issues of defining and identifying giftedness.

Although previous research substantiates the need for modifications in the curriculum for gifted learners (e.g., VanTassel-Baska 2009; Winnebrenner 2004; Merry 2008), nevertheless, gifted learners in most UAE public schools are not provided with curriculum experiences that allow them to reach their full potential (AlGhawi 2017). The common practice in UAE schools to accommodate gifted children is the use of cooperative learning groupings and the use of challenging activities (Elhoweris 2014). However, this may not supply academic benefits for gifted students.

Concluding, so far, only limited research has been conducted to evaluate the effectiveness of enrichment educational models for g/t students in the UAE, which provided a direction for the current study.

3. Rationale for the Study

Findings from several intervention studies in different countries around the world that have used VanTassel-Baska's enrichment model (ICM) in teaching language arts for gifted learners have proven to be successful in promoting verbally gifted students' knowledge, attitudes, motivation, and thinking skills (Feng et al. 2004; Avery and Little 2003; VanTassel-Baska et al. 1996; VanTassel-Baska and Brown 2007, 2009).

As previously mentioned, VanTassel-Baska's (1995) Integrated Curriculum Model (ICM), is one of the most extensively researched curriculum development models in gifted education. The model has demonstrated its ability and effectiveness in providing a learning experience that allows for optimal learning and development of gifted learners in the areas of language arts, science, and social studies (VanTassel-Baska 2009). More specifically, in the area of literacy, the curriculum effectiveness of the ICM model was assessed on US students' literary analysis and interpretations and thinking in persuasive writing by using the four William and Mary language arts units (Feng et al. 2004; Brown et al. 2006). Several other research studies have supported the effectiveness of the use of the ICM model in the subject of language arts (VanTassel-Baska 2015).

However, although this model has proven to be effective, no study has been found that has examined the impact of such reading enrichment programs on UAE/Emirati gifted learners, especially with regards to verbally gifted learners.

Hence, the current study aimed to develop a reading enrichment program based on VanTassel-Baska's ICM model for verbally gifted students in the UAE. The rationale for conducting this study was partially based on the scarcity of research on the area. In addition, the recent trends in critical literacy around the world, focus on critical reading and critical thinking as higher-level functioning traits (Milan 1995), along to invest in high-quality learning as a means for sustaining the renewed economic growth and the well-being of all citizens in the United Arab Emirates, also provided a solid basis for implementing the present study.

Research Questions

The major objective of this study was to examine the impact of a reading enrichment/language arts program, based on VanTassel-Baska's Integrated Curriculum Model, on fourth-grade verbally gifted students in the UAE.

The research questions addressed are the following:

1. To what extent are Emirati verbally gifted fourth-grade students making measurable gains in language arts when working with the differentiated enrichment model?
2. What is the impact of the differentiated enrichment model on the Emirati verbally gifted fourth-grade students' attitudes towards learning?

4. Method

4.1. Participants–Sampling

Forty fourth-grade, nine-year-old students from two schools in the Emirate of Dubai were included in this study (20 males and 20 females). The majority of the participants were UAE nationals (90%), while 10% were from India and Iran.

An experimental group and a control group were formed according to the learning condition. Thus, participants were randomly assigned to either a language arts enrichment condition (n = 20) or a traditional instruction condition (n = 20).

The language arts enrichment program was implemented by two elementary-school Arabic language teachers, who were trained in gifted education, as well as in the basic principle of the newly developed program. Finally, four additional elementary school language teachers and three undergraduate students from the Special Education Department of the United Arab Emirates University (UAEU) were also involved in the implementation of the program and the data analysis procedure as research assistants.

4.2. Development of the Language Arts Enrichment Program

The newly developed enrichment program for the language arts units was created based on VanTassel-Baska's (1995) Integrated Curriculum Model (ICM). More specifically, the language arts enrichment program for the verbally gifted learners included guided critical discussions and advanced organizers for processing, the use of broad themes and concepts, independent research, and interdisciplinary connections.

Since the ICM model emphasizes the use of advanced content, depth, and complexity through student research (VanTassel-Baska 2009), the gifted students attending the language arts enrichment condition, were provided with multiple opportunities to transfer learning from one situation to another, to perform fast processing, and inductive reasoning.

More precisely, the ICM model features three basic, interrelated dimensions. These are (1) Overarching Concepts, (2) Advance Content, and (3) Process-Product.

The Overarching Concepts are based on reading reflections that allow students to develop ideas and themes and determine integrated concepts and ideas originating from different content areas and background knowledge. The Advance Content provides gifted and average students with the opportunity to delve deeper into making synthesis across several content areas, rather than providing shallow ideas. Finally, the Process-Product allows students to explore a topic and conduct research relevant to their selected topic, or engage in a problem-based learning experience (VanTassel-Baska and Little 2011).

In developing the new language arts enrichment program, apart from the ICM's basic dimensions, several other researchers' suggestions were also taken into account. For example, the program involved systematic exposure to high-quality materials (as suggested by Reis et al. 2004), by grouping students based on their reading level and systematic exposure to challenging literature as a means of achieving acceleration, enrichment, as well as critical, creative and inquiry reading experiences (as proposed by Wood 2008).

An additional goal of the program was to establish critical reading and critical thinking enrichment experiences, based on a solid theoretical underpinning for assessing critical reading abilities and critical thinking in the Arabic language for the gifted fourth-grade Emirati students. More specifically, the hierarchical framework followed to design a critical reading assessment, included the levels of structural analysis, rhetoric analysis, social relevance, and holistic evaluation (Applegate et al. 2004; Huijie 2010; Poulson and Wallace 2004). This framework was expected to serve as a reading inventory, based on the theoretical construct of critical reading and critical thinking. The inventory text for guiding students' content writing is comprised of the following components: analyzing paragraphs, discovering meaning, evaluating arguments, and responding to the text. Finally, the overall test focus inventory was accordingly designed to guide the written contents of the enrichment program.

4.3. Procedures and Data Analysis

The two elementary-school Arabic language teachers were trained over a week by the research team on the use of the new language arts enrichment program to implement it with the verbally gifted learners and were provided with professional development opportunities to be able to teach gifted learners.

A quasi-experimental research design was used in the study, which compared the pre-test and post-test performance of the verbally gifted students who attended the language arts enrichment program, as well as with the performance of their gifted peers, who were taught using the mainstream curriculum. This method has been commonly used in previous similar studies examining the outcomes of such programs (e.g., VanTassel-Baska 1995, 2009, 2015).

Participants completed pre-and post-tests on language arts administered by the two teachers in charge. The two teachers, along with the assistance of four other elementary school teachers and three research assistants (UAEU undergraduate students), collected the data. The measure used to assess the participants' critical reading abilities was based on the school formal test that has been used in schools, which is related to the school curriculum. Given the specificities of the Arabic language, the language arts test is commonly used to formally assess the language abilities of primary-aged children in the Emirati countries, as it has been found to provide valid and reliable data. Hence, the reason why we chose to use the language art test for assessing the students' critical reading abilities is that most of the other language testing assessment tools are only available in the English language, which is not the mother language of Emirati students. As several studies have shown, using formal language assessments in Arabic-speaking students might lead to significantly lower scores than proficient English speakers in reading, mathematics, and science and exams, as unfamiliar vocabulary and passive voice constructions may affect the L2 English language learners (ELLs)' comprehension (Abedi and Lord 2001).

Furthermore, as Abedi (2002) argued, the impact of language on ELL assessment was even more obvious in content areas that had higher language demands such as reading, as ELLs may not understand complex questions, may meet unfamiliar vocabulary, and may have a slower reading pace than proficient English speakers (as cited in Ibrahim and Alhosani 2020). Other studies have also supported the use of language art school-based tests for evaluating students' language skills, especially as a means of identifying students who are 'at risk', due to their language proficiency, home language, and immigrant status or due to other demographic characteristics, which is often the case among Emirati students.

Overall, school-based language arts testing has proven to be indifferent to typical, significant, demographic variables including ethnicity and family, whereas their scores also significantly predicted students' later performance in language testing, even after controlling for multiple student and school variables (Goldschmidt and Martinez-Fernandez 2002; Wang et al. 2007).

Data were analyzed using Independent samples t-test statistical analysis to compare and evaluate pre-/post-testing results. Finally, the participating teachers were asked (through an informal short interview) to reflect upon their overall experience in implementing the language enrichment program regarding the benefits they believed it offered both to the verbally gifted students as well as to their own professional development.

To summarize, in this study we used a mixed-method approach. More specifically, quantitative data for the study was collected by using pre-/post-test scores. The qualitative data were collected by using unstructured interviews. The research assistants and the first author asked the participants open-ended questions with respect to their input about the enrichment program. In addition, teachers were also asked about their students' attitudes toward the program. The interviews were conducted within two weeks and followed the flow of a natural conversation. According to the interviewees' availability, the interviews were conducted in the school. These interviews were all recorded using a smartphone application. Each interview lasted for 45–60 min.

For the analysis of the interviews, the process involved a thorough review of the recordings, transcriptions, and interview notes. The coding of procedures of the interviews

followed. The researchers preferred to use Microsoft Word to underline and annotate the participants' answers in order to identify possible themes and introduce visual imagery to make it easy to recognize and relate them. During the process of coding, and analyzing the interview transcripts, several themes emerged.

The researchers used interviews to obtain in-depth data from the teachers about their attitudes toward the implementation of the enrichment program. In addition, teachers were asked about their students' attitudes based on their own observations, which helped the researchers develop a real sense of the students' attitudes toward the enrichment program. Finally, the interview, unlike a survey or a questionnaire, allows the interviewers to modify their questions depending on the respondents' answers.

5. Results

The overall results of this study showed significant pre/post-test student gains and significant differences were revealed between the experimental and control groups in persuasive writing and literary analysis (see Tables S1 and S2 in Supplementary Materials).

To answer the first research question, the means and standard deviations of the post-test were analyzed for both the experimental and control groups. As shown in Table 1 below, the mean score in the pre-test of the control group is 32.30, which is almost equal to their mean score after administering the post-test (32.20), whereas the mean score in the pre-test of the experimental group was 32.70. After implementing the enrichment program, the mean score was 40.30, which indicates higher performance of the experimental group in the post-test.

Table 1. Descriptive Statistics.

Descriptive Statistics					
	N	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Control-Pre-Test	20	20.00	40.00	32.30	5.95
Control-Post-Test	20	21.00	40.00	32.20	6.07
Experimental-Pre-Test	20	25.00	38.00	32.70	3.23
Experimental-PostTest	20	35.00	43.00	40.30	2.62
Valid N (listwise)	20				

An independent samples t-test analysis was run to determine if there were differences in the pre-test and post-test scores of the experimental group after undergoing the enrichment program. The 20 students who attended the enrichment program (M = 40.30, SD = 2.62), compared to the 20 students in the control group (M = 32.20, SD = 6.07), performed significantly better, as shown in the post-test scores, $t(38) = 8.175$, $p < 0.001$ (See Table 2 below).

Table 2. Independent samples t-test (within group).

Independent Samples Test (within Group)								
Groups	**Test**	**N**	**Mean**	**Std. Deviation**	**t-Test**	**df**	**Level of Significance (2-Tailed)**	**Critical t**
Control Group	Pre-Test	20	32.30	5.95	0.053	38	0.958	2.024
	Post-Test	20	32.20	6.07				
Experimental Group	Pre-Test	20	32.70	3.23	−8.175	38	0.000	
	Post-Test	20	40.30	2.62				

An independent samples t-test analysis was run to determine if there were differences in the pre-test scores of the control group versus the experimental group before undergoing the enrichment program. The pre-test scores of the 20 students who attended the enrichment program (M = 40.30, SD = 2.62), compared to the pre-test scores of the 20 students in the control group (M = 32.20, SD = 6.07), were not significantly different (t(38) = 0.264, $p = 0.793$). On the other hand, the 20 students who attended the enrichment program (M = 40.30, SD = 2.62), compared to the 20 students in the control group (M = 32.20, SD = 6.07), performed significantly better, as shown by their post-test scores (t(38) = 5.483, $p < 0.001$ (See Table 3 below). Looking at the post-test mean scores of the control and experimental groups more closely, it was found that the experimental group scored significantly higher (M = 40.30, SD = 2.62) than the control group (M = 32.20, SD = 6.07). According to these results, the enrichment program had a significant positive effect on the students' critical reading abilities, as it was evident that the Emirati fourth-grade, verbally gifted students made significant gains in language arts when working with the differentiated enrichment model. This finding provided a sufficient answer to the second research question of the study.

Table 3. Independent samples t-test (between groups).

Tests	Group	N	Mean	Std. Deviation	t-Test	df	Level of Significance (2-Tailed)	Critical t
Pre-Test	Control	20	32.30	5.95	0.264	38	0.793	2.024
	Experimental	20	32.70	3.23				
Post-Test	Control	20	32.20	9.07	5.483	38	0.000	
	Experimental	20	40.30	2.62				

Concerning the participating students' attitudes toward learning, the participating teachers reported that the experimental group students held positive attitudes towards learning and they were very engaged and interested in the differentiated enrichment model.

Finally, the participating teachers, during the informal interviews, reported that the short training experience they received during this project was considered extremely beneficial regarding their professional development, skills, and confidence in working with verbally gifted students. Moreover, the experience of implementing the language arts enrichment program helped them realize two important facts. First, they were able to witness the significant benefits of effectively providing verbally gifted students with a differentiated and challenging curriculum that met their unique needs and allowed them to reach their high-level potential. Second, they realized that teachers' current training in gifted education is insufficient. More specifically, they claimed that the study revealed some important gaps in teachers' and undergraduate university students' training and professional skills, concerning various aspects of gifted education (e.g., definition/conceptualization of giftedness, assessment methods for identifying gifted learners, and familiarity with the available enrichment programs for teaching g/t students.

6. Discussion

To tap into the UAE's verbally gifted children's potential in language arts, the major objective of this study was to examine the impact of a reading enrichment program, which was developed based on VanTassel-Baska's Integrated Curriculum Model (ICM), on fourth-grade verbally gifted students. In addition, the current study examined the participating verbally gifted students' attitudes towards the enrichment program they attended, as well as the teachers' views and the experiences of those involved in the project.

The overall results of the study indicated that the enrichment program had a significant positive effect on the fourth-grade Emirati verbally gifted students' critical reading abilities,

since the students made significant gains in language arts when working with the differentiated language enrichment model. Moreover, the study revealed that the gifted students held positive attitudes toward the language art enrichment model. This finding confirmed the results of previous studies, which investigated the impact of enrichment reading programs on elementary school students (e.g., Reis et al. 2008). Furthermore, the findings of these studies showed that students in the treatment groups scored statistically significantly higher than those in the control group in reading fluency, reading comprehension, and/or attitudes toward reading, which was also the case in the present study.

The overall benefits of the language arts enrichment model in promoting verbally gifted students' knowledge, motivation, and thinking skills, as evident in our study, agree with the results of several intervention studies in different countries, which evaluated the outcomes of the ICM implementation in teaching language arts to verbally gifted learners (e.g., Brown et al. 2006; Feng et al. 2004; Kim et al. 2012; Gubbins et al. 2002; VanTassel-Baska and Brown 2007).

As mentioned earlier, the newly developed language enrichment program in this study followed the basic dimensions of the ICM, including high-quality reading materials, grouping based on reading levels, and challenging literature, along with literary analysis and persuasive writing activities. Thus, the positive post-test performance of the verbally gifted students who attended the program verified the appropriateness and effectiveness of the aforementioned 'ingredients' that should comprise a successful language arts enrichment program for the verbally gifted. Other researchers have also stressed the importance of including these elements in similar language art programs (e.g., Burkhalter 1995; VanTassel-Baska et al. 2002; Winnebrenner 2004; Wood 2008).

Additionally, it is important to mention that the utilization of the ICM as a baseline for developing our enrichment program was proven a successful choice, thus verifying its developer's statement, according to which the ICM demonstrates a clear design approach since it couples linked subject-based standards with strong elements of differentiation for the gifted learners (VanTassel-Baska 2003). The positive results of our study validate that the particular model and its solid origins in Vygotsky's theory and the ZPD concept, in particular, providing a 'safe' framework for designing challenging, albeit appropriate, learning experiences, which reveal gifted learners' true potential (Burkhalter 1995). Furthermore, the results of the current study confirmed VanTassel-Baska's (2015) claim that the ICM has proven to be a valid basis for motivating both students and teachers to think and learn at higher levels of functioning. The ease in developing our own enrichment model based on the ICM's baselines also validated the coherence in the ICM design and its fidelity of implementation in various educational contexts, also reported by several researchers (e.g., Kim et al. 2012; Feng et al. 2004;VanTassel-Baska et al. 2002). Much earlier, VanTassel-Baska (1995) had already documented the ICM's utility in that several school districts had successfully used the model's baseline to develop their own curricula for gifted students.

The study also provided UAE elementary school teachers with a framework for designing and developing an appropriate curriculum for gifted learners. This finding is in line with previous findings of studies that have confirmed the positive impact of implementing such enrichment programs on teachers' professional development and training skills as regards gifted education provision (AlGhawi 2017; Al Qarni 2010; Feng et al. 2004).

Teachers in this study stated that the whole process of developing and implementing the language arts enrichment program was a very motivating experience, which also provided them with useful skills and tools in terms of identifying, assessing, and planning the next steps of instruction for their gifted students. This finding was also in line with previous research (e.g., VanTassel-Baska 2015). More specifically, the pre- and post-test performance-based assessment procedure helped teachers to better determine their students' general cognitive and linguistic levels and their high ability skills. Similar findings have also been documented in previous studies in the field (e.g., Kim et al. 2012; Feng et al. 2004).

Furthermore, the positive changes in both teachers' and students' attitudes, student motivational response, and the school district change in perspective towards giftedness that is reported in this study, were perceived as resulting from their involvement in the implementation of the language enrichment program. Respective results have been documented in similar studies following the implementation of the ICM science and language arts curricula in gifted education programs (VanTassel-Baska et al. 2000). Teachers in the current study also stated that the whole procedure helped them alter their confidence regarding their competence in differentiating the curriculum to meet the diverse needs of their g/t students. This finding is in line with Stamps' (2004) study results, which revealed that when provided with well-designed programs and with adequate training, teachers feel confident about their skills, and therefore, are eager to alter their teaching methods and habits and differentiate the curriculum to meet the unique needs of 'special' students.

Withregards tothe other language teachers and the undergraduate students who also participated in the study as assistants, they reported an overall positive experience, in terms of the effectiveness of the program on their own professional development. This finding highlighted the need for regularly trained educators to acquire additional skills, knowledge, and training, to be able to support gifted education provisions. Other researchers have also stressed the general lack of professional development of mainstream teachers in terms of effectively supporting gifted learners in various countries (e.g., Hertberg-Davis and M.Callahan 2013), as well as in the UAE (AlGhawi 2017). The assisting teachers and undergraduate students' roles in the project worked using co-teaching and mentoring techniques, which helped them become more skilled, knowledgeable, and aware of working with gifted students. These techniques have been previously proposed as effective forms of professional support (as alternatives to direct training) in special/gifted education contexts (Griffin et al. 2003).

6.1. Implications for Gifted Education Provision in the UAE

Several implications emerge from the current study's findings that have meaning for both researchers and practitioners. The enrichment program used in this study has demonstrated its ability and effectiveness in providing a learning experience that allows for optimal learning and development of gifted learners in the areas of language arts, while it was beneficial for both students and teachers in multiple ways. Hence, the successful implementation of this model provides policymakers in the UAE with empirical data for implementing differentiated enrichment models for teaching gifted learners, which is an innovative aspect of gifted education, and which has not been previously explored in the UAE.

The above conceptualization can be considered an important message for stakeholders in the UAE, in terms of systematically considering the benefits of providing both learners and educators in gifted education with more specific and well-organized enrichment programs. A more systematic organization of such programs in the Emirates would solve several issues often recorded in the everyday gifted education practice. These include the exhibition of frustration and boredom of gifted learners due to the lack of challenging curriculums (Reis et al. 2008; Davis et al. 2011), as well as the frustration of teachers when they are asked to take over the overwhelming task of modifying the curriculum on their own to accommodate a range of diverse students' needs, without having the necessary skills and training (Stamps 2004; AlGhawi 2017).

As previously mentioned, the education system in the UAE has recently emphasized reading as the most important skill students need to master in all academic sections (UAE Innovation Strategy 2015). In addition, the MoE has acknowledged that gifted students have a right to be recognized and catered to within the school (Merry 2008). Furthermore, the recent 'disappointing' results from PISA (OECD 2019), indicated that UAE students fall below the OECD average in reading, along with the worldwide emphasis given on critical and creative reading (Milan 1995). These two matters worked as incentives which positively triggered policymakers to provide verbally gifted students in the Emirati schools with

more challenging and high-ability appropriate curriculums (Colangelo and Davis 2003) and to invest in high-quality learning, in general (Elhoweris 2014). However, recent research revealed various discrepancies and shortcomings regarding the practical application of the official policies and plans suggested by the MoE with regards to gifted education provision in the UAE (AlGhawi 2017; Al-Lawati 2016). More specifically, AlGhawi's study illustrated that gifted education in the UAE is still far from effective, in terms of providing g/t learners with differentiated and challenging curriculums, as well as in terms of teacher training and awareness of the definition, identification, and provision of gifted learners, by both teachers and parents in the UAE.

Therefore, the positive results of the present study regarding the implementation of a language arts enrichment program for verbally gifted students might serve as a useful paradigm for initiating better-organized education provision for the g/t Emirati students. This can either be achieved through the already existing enrichment models (such as the ICM) or through the development of new programs, which will be designed to meet the special educational needs of each school or educational sector. Besides, as Johnsen (2006) stated, gifted education should be seen as a right, rather than a privilege. Finally, as research suggests (Bauwens et al. 1989; Hughes and Murawski 2001; Magiera and Zigmond 2005; Sileo and Garderen 2010), apart from their proven positive effects on gifted education, such innovative enrichment programs could also be beneficial to the general field of Special Educational Needs (SEN).

Summarizing, some of the most important suggestions for improving gifted education provision, deriving from this study, as well as from previous research, can be summarized as follows: Since a distinctive gap has been detected between the national policies for gifted education developed by the UAE MoE and the actual implementation of these policies in the Emirati schools, which causes confusion among teachers, students and parents (AlGhawi 2017), there is a need for a Federal Law to formally acknowledge gifted learners as students with SEN and to safeguard their rights (Elhoweris 2014; Davis and Rimm 2004). This Law should ensure that the official policies and regulations for gifted education become mandatory for both gifted and disabled learners.

A wider and better-organized training plan for pre-service and in-service teachers in gifted education to strengthen human resources in the field, along with constant evaluation practices to improve the quality of gifted education programs (Aljughaiman et al. 2012), should become a part of the current gifted education provision in the UAE. Finally, better practices of raising awareness about giftedness among educators and parents, coupled with more systematic dissemination of policies and successful enrichment programs, should characterize gifted education provision both in the UAE and worldwide.

6.2. Limitations and Future Research

A limitation of the present study was its relatively small sample, which was selected from government schools only in the Emirate of Dubai. This was the mainly due to the difficulty we encountered in finding more, officially diagnosed as verbally gifted, fourth-grade children in the UAE schools. Therefore, in order to be able to generalize our findings, it is recommended that a similar study is conducted in the future that will include a much larger sample of students from various schools across the UAE and from several age groups.

Based on the findings of the current study, some suggestions for future research are the following: It is important to examine the long-term effect of this or similar enrichment programs on language arts and critical reading abilities. Furthermore, future studies should be conducted to provide more empirical data both around the issue of needs assessment and identification of the gifted learners, as well as on the development and implementation of other existing or newly designed enrichment models for meeting the needs of gifted students in various domains, including language arts, science, mathematics, etc.

In addition, the current study could be replicated within a longer period, to re-evaluate the implementation of language art enrichment programs for the verbally gifted at primary schools across the Emirates. Furthermore, since the present study was restricted

to fourth-grade students, it is recommended that future similar studies could focus on the implementation of gifted programs at higher levels of school education (e.g., secondary school).

Another interesting suggestion for future research would be to implement such newly developed enrichment programs for non-gifted students in the UAE schools and assess their possible benefits to these students. Besides, there is evidence from prior research, suggesting that intervention programs with the use of the ICM or other enrichment programs were significantly effective for all students, irrespectively of giftedness (e.g., VanTassel-Baska et al. 2008; Swanson 2006).

In addition, similar studies could be replicated to evaluate the implementation of language-based enrichment programs for verbally gifted learners across different countries. Other studies could be also conducted to evaluate teachers, parents, and students' awareness, perceptions and attitudes towards giftedness as a concept, and/or towards policies and practices of gifted education in the UAE and/or in other countries. Finally, extra research could more systematically evaluate pre-service and in-service teachers' training and professional development in gifted education in the Emirates, to propose more effective ways of enhancing this area.

7. Summary and Conclusions

Gifted children constitute a heterogeneous group of students who manifest high ability and/or talent in several domains. Whether a student is gifted in reading, oral expression, or creative writing, the fact remains that gifted children differ from their peers in the way they think, both quantitatively and qualitatively. Even though giftedness, up to date, has not been captured in a single conceptualization, these differences have led several researchers to argue that gifted children have different educational needs than their counterparts, which cannot be met through curricula designed for their non-gifted peers.

Verbally gifted students, in particular, are those children who typically present higher linguistic abilities than their peers, related to high verbal ability, early reading, advanced vocabulary, and high-level reading comprehension. Therefore, researchers have suggested that, when the gifted reader enters school, instruction must go beyond the traditional basal program, and the focus of reading programs for gifted readers should be on critical and creative reading and thinking.

VanTassel-Baska's (1995) Integrated Curriculum Model (ICM) is one of the most extensively researched curriculum development models in gifted education, which also prioritizes verbal giftedness. The process of instruction and learning included in the ICM curriculum for verbally gifted students emphasizes higher levels of thinking and increased levels of abstractness, while its salient features include accelerated and advanced content, depth, and complexity through abstract concepts, direct study of higher-order thinking processes, interdisciplinary themes, and student research.

Findings from several intervention studies in different countries around the world, that have used the ICM in teaching language arts to gifted learners, provide sufficient evidence that the particular enrichment model was successful in promoting verbally gifted students' knowledge, attitudes, motivation, and thinking skills. Most of these studies agree on the importance of embedding higher-order skills into content and of teaching literary analysis and interpretation, along with persuasive writing, like language arts manifestations of higher-level thinking.

To date, however, no study has been found that investigated the impact of an enrichment program on the UAE verbally gifted children's critical reading abilities. Moreover, recent research revealed various discrepancies and shortcomings regarding the practical application of the official policies and plans suggested by the MoE regarding gifted education provision in the UAE.

Hence, to fulfil UAE gifted children's potential in reading, and to accommodate the educational needs of Emirati verbally gifted students, this study aimed to examine the impact of a reading enrichment program on fourth-grade students' critical reading

abilities. The newly developed program was based on VanTassel-Baska's (2009) Integrated Curriculum Model ICM.

The overall results of the current study indicated that the enrichment program had a significant positive effect on the fourth-grade Emirati verbally gifted students' critical reading abilities since the students made significant gains in language arts when working with the differentiated language enrichment model. Moreover, the study revealed that the gifted students held positive attitudes toward the language arts enrichment model.

In addition, the study provided UAE elementary school teachers with a framework for designing and developing an appropriate curriculum for gifted learners. Hence, there was a positive impact in implementing such an enrichment program on teachers' professional development and training skills regarding gifted education provision. Furthermore, the positive changes in both teachers' and students' attitudes, student motivational response, professional development, and school district change of perspective towards giftedness reported in this study, were perceived as resulting from their involvement in the implementation of the language enrichment program. Additionally, the successful implementation of this model provides policymakers in the UAE with empirical data for implementing differentiated enrichment models for teaching gifted learners, which is an innovative aspect of gifted education, and which has not been previously explored in the UAE.

Concluding, a more systematic organization of such programs in the Emirates would solve several issues often recorded in the everyday gifted education practice. More precisely, a wider and better-organized training plan for pre-service and in-service teachers in gifted education to strengthen human resources in the field, along with constant evaluation practices to improve the quality of gifted education programs should become a part of the current gifted education provision in the UAE. Finally, better practices of raising awareness about giftedness among educators and parents, coupled with more systematic dissemination of policies and successful enrichment programs, should characterize gifted education provision both in the UAE and worldwide.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jintelligence10030068/s1, Table S1: The pre-test scores (out of 45) before the implementation of the enrichment program; Table S2: The post-test scores (out of 45) after the implementation of the enrichment program.

Author Contributions: Conceptualization, H.E. and N.A. (Najwa Alhosani); methodology, N.A. (Najwa Alhosani), H.E. and R.-M.G.B.; software, R.-M.G.B.; validation, H.E.; formal analysis, H.E. and R.-M.G.B.; investigation, N.A. (Negmeldin Alsheikh), E.B. and H.E.; resources, H.E., N.A. (Negmeldin Alsheikh) and R.-M.G.B.; data curation, H.E. and N.A. (Najwa Alhosani); writing—original draft preparation, H.E. and E.B.; writing—review and editing, E.B.; visualization, H.E., N.A. (Najwa Alhosani) and N.A. (Negmeldin Alsheikh); supervision, H.E.; project administration, H.E., N.A. (Negmeldin Alsheikh), N.A. (Negmeldin Alsheikh) and R.-M.G.B.; funding acquisition, H.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the UAEU research office.

Institutional Review Board Statement: The study was conducted in accordance with the UAEU ethical guidelines and approved by the Humanities and Social Sciences Ethics Committee (Application no. ERS_2021_7239).

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 issues. The data are not publicly available due to privacy.

Acknowledgments: We wish to thank the teachers and students who participated in this study.

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

References

Abedi, Jamal. 2002. Standardized Achievement Tests and English Language Learners: Psychometrics Issues. *Educational Assessment* 8: 231–57. [CrossRef]

Abedi, Jamal, and Carol Lord. 2001. The language factor in mathematics tests. *Applied Measurement in Education* 14: 219–34. [CrossRef]
Abu Dhabi Education Council. 2011. Available online: https://www.thenationalnews.com/uae/abudhabi-education-council-made-a-governmentdepartment1.628636 (accessed on 27 April 2022).
Al Qarni, Mohammed A. 2010. *Evaluation of Provisions for Gifted Students in Saudi Arabia*. Wollongong: University of Wollongong.
AlGhawi, Mariam A. 2017. Gifted education in the United Arab Emirates. *Cogent Education* 4: 1368891. Available online: https://www.tandfonline.com/doi/full/10.1080/2331186X.2017.1368891 (accessed on 8 May 2022).
Aljughaiman, Abdullah M., Usama M. A. Ibrahim, and Tayseer M. Khazali. 2012. An evaluation of learning outcomes of summer enrichment gifted programs in Saudi Arabia. Paper presented at the Conference on Creative Education, Shanghai, China, May 20–22.
Al-Lawati. 2016. The Attitudes of GCC Citizens toward the Services Offered to Gifted Students. Available online: https://www.aaas.org/sites/default/files/BTC_Al-Lawati_E.pdf (accessed on 27 April 2022).
Applegate, Mary D., Kathleen B. Quinn, and Anthony J. Applegate. 2004. *Critical Reading Inventory, The: Assessing Students Reading and Thinking and Readers Passages*, 2nd ed. London: Pearson.
Avery, Linda. D, and Catherine Little. 2003. Concept development and learning. In *Content-Based Curriculum for High-Ability Learners*. Edited by Joyce VanTassel-Baska and Catherine Little. Waco: Prufrock Press, pp. 101–24.
Bauwens, Jeanne, Jack J. Hourcade, and Marilyn Friend. 1989. Cooperative teaching: A model for general and special education integration. *Remedial and Special Education* 10: 17–22. [CrossRef]
Beckmann, Else, and Alexander Minnaert. 2018. Non-cognitive Characteristics of Gifted Students With Learning Disabilities: An In-depth Systematic Review. *Frontiers in Psychology* 9: 504. [CrossRef]
Brown, Ann L., and Joseph C. Campione. 1994. Guided discovery in a community of learners. In *Classroom Lessons: Integrating Cognitive Theory and Classroom Practices*. Edited by Kate McGilly. Cambridge: MIT Press, pp. 229–70.
Brown, Elissa, Linda Avery, Joyce Van Tassel-Baska, Bess B. Worley, and Tamra Stambaugh. 2006. A five-state analysis of gifted education policies. Ohio policy study results. *Roeper Review* 29: 11–23. [CrossRef]
Burkhalter, Nancy. 1995. A Vygotsky-based curriculum for teaching persuasive writing in the elementary grades. *Language Arts* 72: 192–96.
Colangelo, Nicholas, and Gary A. Davis. 2003. *Handbook of Gifted Education*, 3rd ed. Boston: Allyn and Bacon, pp. 493–505.
Cross, Tracy L. 2011. *On the Social and Emotional Lives of Gifted Children: Understanding and Guiding Their Development*, 4th ed. Waco: Prufrock Press Inc.
Csikszentmihalyi, Mihaly. 1991. *Flow: The Psychology of Optimal Experience*. New York: Harper Perennial.
Csikszentmihalyi, Mihaly, Kevin R. Rathunde, and Samuel Whalen. 1993. *Talented Teenagers: The Roots of Success and Failure*. New York: Cambridge University Press.
Davis, Gary, and Sylvia Rimm. 2004. *Education of the Gifted and Talented*, 5th ed. Boston: Allyn and Bacon.
Davis, Gary A., Sylvia B. Rimm, and DelSiegle. 2011. *Education of the Gifted and Talented*, 6th ed. Upper Saddler River: Pearson.
Dooley, Cindy. 1993. The Challenge: Meeting the Needs of Gifted Readers. *Reading Teacher* 46: 546–51.
Eddles-Hirsch, Katrina, Wilma Vialle, Karen B. Rogers, and John McCormick. 2010. Just challenge those high-ability learners and they'll be all right: The impact of social context and challenging instruction on the affective development of high-ability students. *Journal of Advanced Academics* 22: 102–68. [CrossRef]
Elhoweris, Hala. 2014. The effect of the label "Giftedness" on the United Arab Emirates pre-service teachers' diagnosis decisions. *International Journal of Education and Research* 2: 515–24.
Feng, Annie Xuemei, Joyce VanTassel-Baska, Chwee Quek, Wenyu Bai, and Barbara O'Neill. 2004. A longitudinal assessment of gifted students' learning using the integrated curriculum model (ICM): Impacts and perceptions of the William and Mary language arts and science curriculum. *Roeper Review* 27: 78–83. [CrossRef]
Ghefli, Andre. 2016. PIRLS 2016 Encyclopedia, The United Arab Emirates, IEA, TIMSS and PIRLS, International Study Center, Lynch School of Education, Boston College. Available online: http://pirls2016.org/ (accessed on 12 May 2022).
Goldschmidt, Pete, and Jose-Felipe Martinez-Fernandez. 2002. *The Relationship among Measures as Empirical Evidence of Validity: Performance Assignments, SAT-9, and High School Exit Exam Performance Incorporating Effects of School Context (CRESST Deliverable)*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Griffin, Cynthia C., Judith A.Winn, Amy Otis-Wilborn, and Karen L. Kilgore. 2003. New teacher induction in special education. Executive Summary. In *Center on Personnel Studies in Special Education*. Gainesville: University of Florida.
Gubbins, E. Jean, Karen L. Westberg, Sally M. Reis, Susan T. Dinnocenti, Carol L. Tieso, Lisa M. Muller, Sunghee Park, Linda J. Emerick, Lori R. Maxfield, and Deborah E. Burns. 2002. *Implementing a Professional Development Model Using Gifted Education Strategies with All Students*. (Report RM02172). Storrs: University of Connecticut, National Research Center on the Gifted and Talented.
Hertberg-Davis, Holly L., and Carolyn M.Callahan. 2013. Introduction. In *Fundamentals of Gifted Education*. Edited by Holly L. Hertberg-Davis and Carolyn M. Callahan. New York: Routledge, pp. 1–10.
Hughes, Claire E., and Wendy A. Murawski. 2001. Lessons from another field: Applying coteaching strategies to gifted education. *Gifted Child Quarterly* 45: 195–204. [CrossRef]
Huijie, Lin. 2010. Developing a hierarchical framework of critical reading proficiency. *Chinese Journal of Applied Linguistics* 33: 40–54.
Ibrahim, Ali, and Najwa Alhosani. 2020. Impact of language and curriculum on student international exam performances in the United Arab Emirates. *Cogent Education* 7: 1808284. [CrossRef]

Jiboye, Temitope Favour, Gafar Olaide Salaudeen, Oluwabunmi Opeyemi Adejumo, and David Olaniyi Aikomo. 2019. Mental ability, Self-esteem and Learning Styles as Correlate of Creativity among High Achieving Secondary School Students in Oyo State. *International Journal of Innovation, Creativity and Change* 4: 24–43. Available online: https://www.ijicc.net/images/vol4iss4/JIBOYE_et_al_2019.pdf (accessed on 19 January 2022).

Johnsen, Susan K. 2006. New national standards for teachers of gifted and talented students. *Tempo* 26: 26–31.

Kim, Kyung Hee, Joyce VanTassel-Baska, Bruce A. Bracken, Annie Feng, Tamra Stambaugh, and Lori Bland. 2012. Project Clarion: Three years of science instruction in title I schools among K–third-grade students. *Research in Science Education* 42: 813–29. [CrossRef]

Knowledge and Human Development Authority. 2011. Available online: www.khda.gov.ae (accessed on 28 March 2022).

Magiera, Kathleen, and Naomi Zigmond. 2005. Co-teaching in middle school classrooms under routine conditions: Does the instructional experience differ for students with disabilities in co-taught and solo-taught classes? *Learning Disabilities Research and Practice* 20: 79–85. [CrossRef]

Maker, C. June. 2005. *The DISCOVER Project: Improving Assessment and Curriculum for Diverse Gifted Learners (RM05206)*. Storrs: The National Research Center on the Gifted and Talented.

Martin, Michael O., Ina V. S. Mullis, and Ann M. Kennedy, eds. 2007. Progress in International Reading Literacy Study (PIRLS): PIRLS 2006 Technical Report, International Association for the Evaluation of Educational Achievement, TIMSS and PIRLS. Available online: https://files.eric.ed.gov/fulltext/ED499436.pdf (accessed on 7 April 2022).

McLaughlin, Milbrey W., and Joan E. Talbert. 1993. *Teaching for Understanding: Challenges for Policy and Practice*. San Francisco: Jossey-Bass.

Merry, Michael S. 2008. Educational justice and the gifted. *School Field* 6: 47–70. [CrossRef]

Milan, Deanne. 1995. *Developing Reading Skills*. New York: McGraw-Hill, Inc.

Monks, Franz J., and Emanuel J. Mason. 2000. Developmental psychology and giftedness: Theories and research. In *International Handbook of Giftedness and Talent*. Edited by Kurt A. Heller, Franz J. Monks, Robert J. Sternberg and Rena F. Subotnik. Oxford: Elsevier Science, pp. 587–94.

National Association for Gifted Children. 2010. Redefining Giftedness for a New Century: Shifting the Paradigm. Available online: https://www.nagc.org/sites/default/files/Position%20Statement/Redefining%20Giftedness%20for%20a%20New%20Century.pdf (accessed on 9 February 2022).

Newstead, Stephen E., and Peter C. Wason, eds. 1995. *Perspectives on Thinking and Reasoning: Essays in Honor of Peter Wason*. Hillsdale: Erlbaum.

OECD. 2019. *PISA 2018 Assessment and Analytical Framework*. Paris: PISA, OECD Publishing. [CrossRef]

OECD. 2021. *21st-Century Readers: Developing Literacy Skills in a Digital World*. Paris: PISA, OECD Publishing. [CrossRef]

Paul, Catherine A. 2015. *The Federal Elementary and Secondary Education Act*. Available online: https://socialwelfare.library.vcu.edu/programs/education/elementary-and-secondary-education-act-of-1965/ (accessed on 9 February 2022).

Poulson, Louise, and Mike Wallace. 2004. Designing and writing about research: Developing a critical frame of mind. In *Learning to Read Critically in Teaching and Learning*. Edited by Louise Poulson and Mike Wallace. London: Sage Publication.

Reis, Sally M., E. Jean Gubbins, Christine J. Briggs, Fredric J. Schreiber, Susannah Richards, Joan K. Jacobs, Rebecca D. Eckert, and Joseph S. Renzulli. 2004. Reading instruction for talented readers: Case studies documenting few opportunities for continuous progress. *Gifted Child Quarterly* 48: 315–38. [CrossRef]

Reis, Sally M., Elizabeth A. Fogarty, Rebecca D. Eckert, and Lisa M. Muller. 2008. *Schoolwide Enrichment Model Reading Framework*. New York: Prufrock Press.

Renzulli, Joseph S. 1977. *The Enrichment Triad Model: A Guide for Developing Defensible Programs for the Gifted and Talented*. Mansfield Center: Creative Learning Press.

Renzulli, Joseph S. 1978. What makes giftedness? Re-examining a definition. *Phi Delta Kappan* 60: 180–84.

Renzulli, Joseph S. 2005. The Three-Ring Conception of Giftedness: A Developmental Model for Promoting Creative Productivity. In *Conceptions of Giftedness*. Edited by Robert J. Sternberg and Janet E. Davidson. New York: Cambridge University Press, pp. 246–79.

Renzulli, Joseph S., Sally M. Reis, and Linda Smith. 1981. The Revolving-Door Model: A new way of identifying the gifted. *Phi Delta Kappan* 62: 648–49.

Scruggs, Thomas E., and Sanford J. Cohn. 1983. Learning Characteristics of Verbally Gifted Students. *Gifted Child Quarterly* 27: 169–72. [CrossRef]

Sileo, Jane M., and Delindavan Garderen. 2010. Creating optimal opportunities to learn mathematics: Blending co-teaching structures with research-based practices. *Teaching Exceptional Children* 42: 14–21. [CrossRef]

Stamps, Lisa S. 2004. On Teaching Gifted Students: The Effectiveness of Curriculum Compacting in First Grade Classrooms. *Roeper Review* 27: 31–41. [CrossRef]

Swanson, Julie Dingle. 2006. Breaking Through Assumptions About Low-Income, Minority Gifted Students. *Gifted Child Quarterly* 50: 11–25.

UAE Innovation Strategy. 2015. Available online: https://www.moei.gov.ae/assets/download/1d2d6460/National%20Innovation%20Strategy.pdf.aspx (accessed on 10 May 2022).

UAE Vision. 2010. United in Ambition and Determination. Available online: https://www.vision2021.ae/en/uae-vision (accessed on 26 January 2022).
VanTassel-Baska, Joyce. 1995. The development of talent through the curriculum. *Roeper Review* 18: 98–102. [CrossRef]
VanTassel-Baska, Joyce. 2003. *Curriculum Planning and Instructional Design for Gifted Learners*, 2nd ed. Denver: Love.
VanTassel-Baska, Joyce. 2008. *Assessment for Gifted Students*. Waco: Prufrock Press.
VanTassel-Baska, Joyce. 2009. Affective curriculum and instruction for gifted learners. In *The Critical Issues in Equity and Excellence in Gifted Education Series. Social-Emotional Curriculum with Gifted and Talented Students*. Edited by Joyce L. VanTassel-Baska, Tracy L. Cross and F. Richard Olenchak. Austin: Prufrock Press Inc., pp. 113–32.
VanTassel-Baska, Joyce. 2015. Differentiation in action: The Integrated Curriculum Model. *Revista de Educacion* 368: 225–44. [CrossRef]
VanTassel-Baska, Joyce, and Catherine Little. 2011. *Content-Based Curriculum for the Gifted*. Waco: Prufrock Press.
VanTassel-Baska, Joyce, and Elissa F. Brown. 2007. Toward Best Practice: An Analysis of the Efficacy of Curriculum Models in Gifted Education. *Gifted Child Quarterly* 51: 342–58. [CrossRef]
VanTassel-Baska, Joyce, and Elissa F. Brown. 2009. An analysis of gifted education curriculum models. In *Methods and Materials for Teaching the Gifted*. Edited by Frances A. Karnes and Suzanne M. Bean. Austin: Prufrock Press Inc., pp. 75–106.
VanTassel-Baska, Joyce, and Tamra Stambaugh. 2006. *Comprehensive Curriculum for the Gifted*. Boston: Pearson.
VanTassel-Baska, Joyce, Annie Xuemei Feng, Elissa F. Brown, Bruce Bracken, Tamra Stambaugh, Heather French, Susan McGowan, Bess Worley, Chwee Quek, and Wenyu Bai. 2008. A study of differentiated instructional change over three years. *Gifted Child Quarterly* 52: 297–312. [CrossRef]
VanTassel-Baska, Joyce, Bruce Bracken, Annie Feng, and Elissa F. Brown. 2009. A longitudinal study of reading comprehension and reasoning ability of students in elementary Title I schools. *Journal for the Education of the Gifted* 33: 7–37. [CrossRef]
VanTassel-Baska, Joyce, Dana T. Johnson, Claire E. Hughes, and Linda Neal Boyce. 1996. A Study of Language Arts Curriculum Effectiveness with Gifted Learners. *Journal for the Education of the Gifted* 19: 461–80. [CrossRef]
VanTassel-Baska, Joyce, Li Zuo, Linda D. Avery, and Catherine A. Little. 2002. A curriculum study of gifted student learning in the language arts. *Gifted Child Quarterly* 46: 30–44. [CrossRef]
VanTassel-Baska, Joyce, Linda D. Avery, Claire E. Hughes, and Catherine A. Little. 2000. An evaluation of the implementation of curriculum innovation: The impact of William and Mary units on schools. *Journal for the Education of the Gifted* 23: 244–72. [CrossRef]
Vye, Nancy J., Susan R. Goldman, James F. Voss, Cindy Hmelo, and Susan Williams. 1998. Complex mathematical problem solving by individuals and dyads. *Cognition and Instruction* 15: 435–84. [CrossRef]
Vygotsky, Lev S. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Cambridge: Harvard University Press.
Wang, Jia, David Niemi, and Haiwen Wang. 2007. *Haiwen Predictive Validity of an English Language Arts Performance Assessment, CRESST REPORT 729*; Los Angeles: University of California, pp. 1–27. Available online: https://files.eric.ed.gov/fulltext/ED499438.pdf (accessed on 2 May 2022).
Winnebrenner, Susan. 2004. *Teaching Gifted Kids in the Regular Classroom: Strategies and Techniques Every Teacher Can Use to Meet the Academic Needs of the Gifted and Talented*. Michigan: Free Spirit Publishing.
Wood, Patricia F. 2008. Reading Instruction With Gifted and Talented Readers: A Series of Unfortunate Events or a Sequence of Auspicious Results? *Gifted Child Today* 31: 16–25. [CrossRef]
Wynn, Karen. 1990. Children's understanding of counting. *Cognition* 36: 155–93. [CrossRef]

Journal of
Intelligence

MDPI

Article

Linguistic Influences on Cognitive Test Performance: Examinee Characteristics Are More Important than Test Characteristics

Damien C. Cormier [1,*], Okan Bulut [2,*], Kevin S. McGrew [3] and Kathleen Kennedy [1]

[1] Department of Educational Psychology, University of Alberta, Edmonton, AB T6G 2G5, Canada; kk4@ualberta.ca
[2] Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB T6G 2G5, Canada
[3] Institute for Community Integration, University of Minnesota and Institute for Applied Psychometrics, St. Joseph, MN 56374, USA; iqmcgrew@gmail.com
* Correspondence: dcormier@ualberta.ca (D.C.C.); bulut@ualberta.ca (O.B.)

Abstract: Consideration of the influence of English language skills during testing is an understandable requirement for fair and valid cognitive test interpretation. Several professional standards and expert recommendations exist to guide psychologists as they attempt to engage in best practices when assessing English learners (ELs). Nonetheless, relatively few evidence-based recommendations for practice have been specified for psychologists. To address this issue, we used a mixed-effects modeling approach to examine the influences of test characteristics (i.e., test directions) and examinee characteristics (i.e., expressive and receptive language abilities) on cognitive test performance. Our results suggest that language abilities appear to have a significant influence on cognitive test performance, whereas test characteristics do not influence performance, after accounting for language abilities. Implications for practice include the assessment of expressive and receptive language abilities of EL students prior to administering, scoring, and interpreting cognitive test scores.

Keywords: psychoeducational assessment; cognitive abilities; language abilities; school psychology; clinical psychology

Citation: Cormier, Damien C., Okan Bulut, Kevin S. McGrew, and Kathleen Kennedy. 2022. Linguistic Influences on Cognitive Test Performance: Examinee Characteristics Are More Important than Test Characteristics. *Journal of Intelligence* 10: 8. https://doi.org/10.3390/jintelligence10010008

Received: 8 September 2021
Accepted: 24 January 2022
Published: 27 January 2022

1. Introduction

It is widely accepted that diagnostic decisions need to be based on strong assessment practices. There are several indicators suggesting that the diagnostic process can be challenging when a standardized, norm-referenced test is used to assess students from diverse backgrounds. One is the disproportionate number of students from diverse backgrounds qualifying for special education services under a variety of disability categories (Harry and Klingner 2014). This problem has been described as being "among the most longstanding and intransigent issues in the field" (Skiba et al. 2008, p. 264). Although multiple factors contribute to the diagnostic process, some researchers have focused their attention on a type of test (e.g., cognitive measures) used to inform their diagnostic decisions (e.g., Cormier et al. 2014; Styck and Watkins 2013). However, even as researchers focus on one type of test, there is a myriad of potential contributing factors that they examine. For example, some researchers have investigated potential differences in the linguistic environments of students from diverse backgrounds compared to those of the majority, which are better represented in a test's normative sample (Ortiz et al. 2012). Regardless of their focus, one of the limitations of many of these studies is that the group of interest is defined as culturally and linguistically diverse students, which is very broad. More recently, Ortiz (2019) has argued for a narrower focus on English learners (EL), so recommendations for practice can be applied more appropriately by considering a single attribute, such as their language abilities.

English learner (EL) has frequently been used as the primary descriptor for the group of students who are likely going to struggle to demonstrate their true abilities on a standardized measure that is normed using an English-speaking population (Ortiz 2019). EL specifically refers to students who are non-native English speakers with acquired proficiency in a second language (Ortiz 2019). At times, the term EL is confused with bilingual. This confusion is problematic because the word bilingual does not necessarily imply that English is the second language acquired by the student. EL is more appropriate given that the students within this group can fall at various points along the English proficiency continuum (Ortiz 2019). Furthermore, any given level of English proficiency does not negate the fact that these students are either continuing to develop their proficiency skills in English or are working towards doing so (Ortiz 2019). To improve accuracy in diagnostic decisions, psychologists and other professionals conducting assessments must take a comprehensive approach to understand both the examinee and test characteristics that may influence test performance for EL students when they are administered standardized, norm-referenced tests.

1.1. Assessing the Abilities of EL Students

The valid assessment of EL students often poses a significant challenge for psychologists. Although a number of broad and specific professional standards exist, best practices are still unclear. For example, Standard 9.9 from the *Standards for Psychological and Educational Testing* (AERA et al. 2014) emphasizes the requirement of having a "sound rationale and empirical evidence, when possible, for concluding that [. . .] the validity of interpretations based on the scores will not be compromised" (p. 144). The *Standards* further stress that when a test is administered to an EL student, "the test user should investigate the validity of the score interpretations for test-takers with limited proficiency in the language of the test" (p. 145). The commentary associated with this specific standard highlights the responsibility of the psychologist in ensuring that language proficiency is minimized as a factor in the interpretation of the test scores.

If practicing psychologists were to consult the available evidence, they would need to weigh multiple variables and attempt to infer the extent to which a particular student's assessment data is influenced by such variables. To date, some studies have suggested varying degrees of observed score attenuation for linguistically diverse students (e.g., Cormier et al. 2014; Kranzler et al. 2010; Tychanska 2009). Other studies have reported that the rate of differentiation between groups is no better than chance (Styck and Watkins 2013). Regardless of the underlying causal variable, these studies only suggest that practitioners should be cautious when interpreting test results. Thus, the extant literature does not provide specific recommendations or strategies for generating or testing hypotheses that would help practitioners better understand which test or examinee characteristics influence observed test score performance. Nonetheless, the two key areas initially identified and described by Flanagan et al. (2007)—test characteristics and examinee characteristics—continue to be focal areas of interest, as researchers attempt to develop approaches to assessing EL students that are in line with best practices.

1.1.1. Test Characteristics

For decades, the content, administration, scoring, and interpretation of tests—essentially, tests' psychometric properties—were highly criticized as producing biased scores for EL students (see Reynolds 2000, for a comprehensive review). Years of research produced arguments (e.g., Helms 1992) and counterarguments (e.g., Brown et al. 1999; Jensen 1980) with respect to this controversy. Although there was some evidence of bias in the measures used in previous decades, test developers appear to have taken note of the limitations of previous editions and, as a result, contemporary measures have significantly reduced the psychometric biases between groups (Reynolds 2000).

Despite these advances within the test development process, researchers have continued to investigate potential issues. For example, Cormier et al. (2011) were the first to

quantify the linguistic demand of the directions associated with the administration of a standardized measure. Their goal was to examine the relative influence of this variable across a measure's individual tests (i.e., subtests); if there was considerable variability among a measure's individual tests, then this may be a meaningful variable to consider as practitioners selected their battery when assessing EL students. They accomplished this goal by using an approach suggested by Dr. John "Jack" Carroll, a prominent scholar who was not only a major contributor to the initial Cattell-Horn-Carroll (CHC) Theory of Intelligence, but who also had a passion for psycholinguistics. The methodology involved using text readability formulas to approximate the linguistic demand of test directions required with the 20 tests from the Woodcock-Johnson Tests of Cognitive Abilities, Third Edition (WJ III; Woodcock et al. 2001). A subsequent study (Cormier et al. 2016) investigated the linguistic demand of test directions across two editions of the same cognitive battery—the WJ III and the Woodcock-Johnson Tests of Cognitive Abilities, Fourth Edition (WJ IV COG; Schrank et al. 2014a). Eventually, Cormier et al. (2018) identified several test outliers across commonly used cognitive test batteries by applying the methodology used by Cormier et al. (2011). Although this series of studies appears to have produced meaningful recommendations to practitioners with respect to test selection, the extent to which test directions have a significant influence on the actual performance of examinees remains unknown.

1.1.2. Examinee Characteristics

Standardized measures are constructed based on the presumption that the students who are assessed using the measures possess a normative level of English proficiency. (Flanagan et al. 2007). Under this presumption, the average examinee would be able to understand test directions, produce verbal responses, or otherwise use their English language skills at a level that is consistent with their peers (Flanagan et al. 2007). However, as noted by Ortiz (2019), there is likely to be a continuum of language levels within an EL population. Unfortunately, researchers have not focused on the level of English language proficiency as a way of differentiating performance on standardized measures. As a result, much of the research completed to date only provides information on general group-level trends that may or may not be observed by practitioners *after* they have administered a battery of standardized measures. One of the potential reasons for the lack of consistency may be related to the population of interest being defined as culturally *and* linguistically diverse, instead of focusing on a specific, measurable student characteristic, such as English language proficiency. As a result, practitioners may be reluctant to incorporate additional measures into their assessment batteries when testing EL students because: (a) there is no clear definition of the type of student these recommendations would apply to; and (b) the evidence regarding the validity of patterns of performance on standardized measures is, at best, still mixed (e.g., Styck and Watkins 2013). Thus, there appears to be a need to examine the influence of both test and examinee characteristics together to better understand the impact of these characteristics on test performance when standardized measures are used.

1.2. Current Study

A review of the literature has led to the conclusion that researchers have, perhaps, overlooked the potential of considering examinee characteristics as they attempt to produce research with empirical recommendations for the assessment of EL students. Moreover, the data produced by Cormier et al. (2018) provide a quantification of the linguistic demands of many commonly used measures of cognitive abilities. Taken together, it is now possible to investigate test and examinee characteristics together, to understand their relative contributions to performance on standardized measures. Consequently, we sought to answer the following research question: What are the relative contributions of test characteristics (i.e., the linguistic demand of test directions) and examinee characteristics (i.e., oral English language skills) to performance on a standardized measure of cognitive abilities?

2. Materials and Methods

2.1. Sample

The normative sample for the Woodcock-Johnson IV (Schrank et al. 2014b) was the primary source of data for this study. The Woodcock-Johnson IV is a battery of 51 tests that includes the WJ IV COG, the Woodcock-Johnson Tests of Academic Achievement (WJ IV ACH; Schrank et al. 2014c), and the Woodcock-Johnson Tests of Oral Language (WJ IV OL; Schrank et al. 2014d). The stratified sampling design included variables such as region, sex, country of birth, race, ethnicity, community type, parent education, and educational attainment. For the purpose of this study, we used the school-age sub-sample, which resulted in a sample size of 4212 students.

2.2. Measures

The WJ IV COG is comprised of 18 individual tests; 10 for the standard battery and eight for the extended battery (see Table 1 for a list of the WJ IV COG tests). Four of the WJ IV COG tests—Oral Vocabulary, Phonological Processing, Visualization, and General Information—contain subtests (see Table 1). The 18 tests of cognitive abilities were developed for the purpose of "measuring general intellectual ability, broad and narrow cognitive abilities, academic domain-specific aptitudes, and related aspects of cognitive functioning" (McGrew et al. 2014, p. 8). Seven tests from the WJ IV COG are used to generate a General Intellectual Ability (GIA) score for assessing the higher-order psychometric *g* construct from the CHC Theory of Intelligence. Overall, the WJ IV COG demonstrates excellent technical adequacy (see Reynolds and Niileksela 2015 for a comprehensive review of the WJ IV COG). For this study, we used individual test and subtest scores, which also demonstrate strong psychometric properties. For example, the internal consistency coefficients for the WJ IV COG individual tests range "from 0.74 to 0.97, with a median reliability of 0.89" (Reynolds and Niileksela 2015, pp. 387–88). The validity evidence is also strong and was described as "strikingly comprehensive" (p. 388).

The WJ IV OL is comprised of 12 tests (see Table 1 for a list of the WJ IV OL tests). The stated purpose of the WJ IV OL is to measure "oral language ability and listening comprehension (in English or Spanish), oral expression, and two important cognitive-linguistic abilities: phonetic coding and speed of lexical access" (McGrew et al. 2014, p. 10). For this study, we used the Oral Expression and Listening Comprehension cluster scores that are administered in English only (see Table 1 for test details). The median reliability coefficients for the Oral Expression and Listening Comprehension clusters are 0.89 and 0.90, respectively (McGrew et al. 2014, p. 291).

As noted previously, the study completed by Cormier and colleagues produced quantitative values for the relative influence of the linguistic demand of test directions across cognitive assessment batteries. Their investigation included test directions from four norm-referenced measures of cognitive abilities: the Cognitive Assessment System, Second Edition, the Kaufman Assessment Battery for Children, Second Edition, the Wechsler Intelligence Scale for Children, Fifth Edition, and the WJ IV COG. A total of 99 individual tests and subtests from these four measures were included in their analyses. Principal component analyses produced values ranging from -0.96 to 5.37 for standard test directions and values ranging from -0.57 to 8.39 for supplementary test directions.

The relative linguistic demand values for the WJ IV COG used for analysis were taken from their results. The analysis was completed at the subtest level, when applicable because unique directions are provided for each subtest. The only exception is the General Information subtests, *Where* and *When*, which was represented at the test level in the analysis completed by Cormier and colleagues, presumably due to their data coding rules. Thus, the final tests and subtests from the WJ IV COG used for the purposes of this study are listed in Table 1.

Table 1. List of WJ IV Tests of Cognitive Abilities and WJ IV Tests of Oral Language.

WJ IV Tests of Cognitive Abilities—Subtest	WJ IV Tests of Oral Language (English Tests Only)
Oral Vocabulary—Synonyms	Picture Vocabulary [2]
Oral Vocabulary—Antonyms	Oral Comprehension [3]
Number Series	Segmentation
Verbal Attention	Rapid Picture Naming
Letter-Pattern Matching	Sentence Repetition [2]
Phonological Processing—Word Access	Understanding Directions [3]
Phonological Processing—Word Fluency	Sound Blending
Phonological Processing—Substitution	Retrieval Fluency
Story Recall	Sound Awareness
Visualization—Spatial Relations	
Visualization—Block Rotation	
General Information [1]	
Concept Formation	
Numbers Reversed	
Number-Pattern Matching	
Nonword Repetition	
Visual-Auditory Learning	
Picture Recognition	
Analysis-Synthesis	
Object-Number Sequencing	
Pair Cancellation	
Memory for Words	

[1] The General Information test score is produced from the General Information—What and General Information—Where subtests. The subtests are not listed in the table because the analysis was completed at the test level for General Information; [2] The Picture Vocabulary and Sentence Repetition tests produce the Oral Expression cluster score; [3] The Oral Comprehension and Understanding Directions tests produce the Listening Comprehension cluster score.

2.3. Procedure

To determine the relative contributions of the variables of interest, we employed a mixed-effects modeling approach. Unlike linear regression models that can only estimate fixed effects for independent variables, mixed-effects models can incorporate both fixed and random effects within a multilevel structure. In the context of mixed-effects modeling, random effects refer to parameters of independent variables that represent a random sample of variables drawn from a population with mean μ and standard deviation σ. A typical mixed-effects model can be expressed in matrix form as follows:

$$y_i = \mathbf{X_i}\beta + \mathbf{Z_i}b_i + \varepsilon_i, \quad (1)$$

where y_i is the vector of the dependent variable y for group i (i = 1, 2, 3, . . . , K) where individual observations (level 1) are nested within the group (level 2), $\mathbf{X_i}$ is the matrix of the fixed-effect predictors for group i, β is the vector of fixed-effect coefficients which are the same for all groups, $\mathbf{Z_i}$ is the matrix of the random effects for group i, b_i is the vector of random-effect coefficients for group i, and ε_i is the error (i.e., residual) term for individual observations in group i. In Equation (1), both b_i and ε_i are assumed to be normally distributed.

In this study, we assume a multilevel structure where students (level 1) are nested within the tests of the WJ IV COG (level 2). That is, the tests from the WJ IV COG represent a sample of tests drawn from the population of all possible cognitive tests. To analyze the effects of the student- and test-related predictors, we developed six mixed-effects models (i.e., Models 1 to 6). We opted to use *W* scale scores from the WJ IV COG and the WJ IV OL, which are "a direct transformation of the Rasch logit scale" (McGrew et al. 2014, p. 46). The *W* scores are centered on a value of 500, which eliminates the possibility of negative ability scores. Model 1 aimed to demonstrate the variation in students' *W* scores across

the tests, and thus it did not include any fixed-effect predictors. Model 2 included age and the GIA scores as fixed-effect predictors to account for the variation in the test scores due to students' cognitive development and individual differences in general intellectual functioning, respectively. Model 3 included age and GIA scores at the student level and the linguistic demand of test directions and its interaction with age at the test level as fixed-effect predictors. The interaction between age and the linguistic demand of test directions was included to account for a potentially greater effect on younger students than on older students. This effect assumes that older students are typically less likely to have difficulty understanding the test directions compared to younger students. Models 4 and 5 included students' oral expression and listening comprehension scores from the WJ IV OL as additional predictors, respectively. Finally, Model 6 included all the predictors (i.e., age, GIA, linguistic demand of test directions, and its interaction with age, oral expression, and listening comprehension) to predict the *W* scores. The mixed-effects models were estimated using the *lme4* package (Bates et al. 2015) in R (R Core Team 2021). The models were evaluated in terms of the statistical significance of the fixed-effect predictors, as well as the change in the amount of variance explained by the models.

3. Results

The results of the mixed-effects models are summarized in Tables 2 and 3. The variance estimates for Model 1 indicated that there was considerable variation in the students' scores at both levels 1 and 2. The random effect estimates in Figure 1 essentially represent the differences between the average scores for the tests and the average score across all tests (i.e., vertical, dashed line). In addition, the WJ IV COG tests on the *y*-axis are sorted in descending order by the linguistic demand of the test directions. Relative to the overall average score, the average score for Memory for Words was the smallest, whereas the average score for Letter-Pattern Matching was the highest. Figure 1 shows that although the Gc tests (i.e., Comprehension-Knowledge) seem to have higher linguistic demand than the remaining tests, there is no systematic pattern in terms of the relationship between the average scores and the linguistic demand of their test directions. Figure 1 also indicates that the WJ IV COG tests differ regarding students' performance on the tests. Thus, test- and student-level predictors can be used to further explain the variation among the tests.

The first model with fixed-effect predictors was Model 2. Table 2 shows that both the GIA scores and age were significant, positive predictors of student performance on individual cognitive tests. This result was anticipated given that all the abilities measured by the WJ IV COG are expected to continue to develop at an accelerated rate within the age range of the sample used for this study (see McGrew et al. 2014, p. 137). Similarly, the GIA score is expected to have a positive relationship with individual cognitive tests because it is the composite of the theoretical constructs (i.e., first-order factors) underlying the individual tests in the WJ IV COG. The GIA score is interpreted as a robust measure of statistical or psychometric *g*.

It should be noted that there are renewed debates regarding what intelligence test global composite scores (e.g., GIA, Wechsler Full-Scale IQ) represent. Briefly, the finding of a statistical or psychometric *g* factor is one of the most robust findings over the last 100+ years in psychology (Wasserman 2019). However, recent statistical and theoretical advances using network science methods (viz., psychometric network analysis; PNA) have suggested that psychometric *g* is only a necessary mathematical convenience and a statistical abstraction. The reification of the *g* factor in psychometrics is due, in large part, to the conflation of psychometric and theoretical *g* and has contributed to the theory crises in intelligence research (Fried 2020). Researchers using contemporary cognitive theories of intelligence (e.g., dynamic mutualism; process overlap theory; wired intelligence) have shown valid alternative non-latent trait common cause factor explanations of the positive manifold of intelligence tests. Furthermore, the previously mentioned period of 100+ years of research regarding general intelligence has also robustly demonstrated that there is yet no known biological or cognitive process theoretical basis of psychometric *g* (Barbey 2018;

Detterman et al. 2016; Kovacs and Conway 2019; van der Maas et al. 2019; Protzko and Colom 2021). Therefore, in this paper, the GIA score is interpreted to reflect an emergent property that is a pragmatic statistical proxy for psychometric *g*, and not a theoretical individual differences latent trait characteristic of people. This conceptualization of the GIA score is similar to the emergent property of an engine's horsepower: an emergent property index that summarizes the efficiency of the complex interaction of engine components (i.e., interacting brain networks), in the absence of a "horsepower" component or factor (i.e., theoretical or psychological *g*). Given that this debate is still unresolved after 100+ years of research, the GIA cluster was left in the analysis to acknowledge the possibility that theoretical or psychological *g* may exist and to recognize the strong pragmatic predictive powers of such a general proxy. Leaving GIA out of the analysis, which would reflect a strong "there is no *g*" position (McGrew et al. 2022), reflects the authors' recognition that many intelligence scholars still maintain a belief in the existence of an elusive underlying biological brain-based common cause mechanism. More importantly, the inclusion of GIA recognizes the overwhelming evidence of the robust pragmatic predictive power of psychometric *g*. As such, the inclusion of age and the GIA scores in the model allowed us to control for their effects when including additional variables. The inclusion of age and the GIA scores also explained a large amount of variance at the student level (level 1), reducing the level-1 unexplained variance from 573.39 to 265.72 (i.e., 46% reduction).

Table 2. A Summary of the Results from the Mixed-Effects Models.

		Variances		Fixed Effects		
Model	**Predictors**	**Level 1**	**Level 2**	β	t	p
1	Intercept	573.39	25.12	502.10	469	<0.001
2	Intercept	265.72	25.14	45.67	19.88	<0.001
	Age			0.02	9.83	<0.001
	GIA			0.90	203.84	<0.001
3	Intercept	265.68	25.13	45.59	19.79	<0.001
	Age			0.02	10.15	<0.001
	GIA			0.90	203.85	<0.001
	Test directions			0.53	0.47	0.643
	Test directions × Age			−0.01	−3.61	<0.001
4	Intercept	265.28	25.13	41.56	17.86	<0.001
	Age			0.01	8.41	<0.001
	GIA			0.86	145.88	<0.001
	Test directions			0.53	0.47	0.643
	Test directions × Age			−0.01	−3.61	<0.001
	Oral expression			0.05	11.79	<0.001
5	Intercept	265.24	25.13	33.46	13.37	<0.001
	Age			0.02	9.16	<0.001
	GIA			0.84	128.74	<0.001
	Test directions			0.53	0.47	0.643
	Test directions × Age			−0.01	−3.61	<0.001
	Listening comprehension			0.08	12.38	<0.001
6	Intercept	265.12	25.13	34.53	13.77	<0.001
	Age			0.01	8.34	<0.001
	GIA			0.83	123.01	<0.001
	Test directions			0.53	0.47	0.643
	Test directions × Age			−0.01	−3.61	<0.001
	Oral expression			0.03	6.52	<0.001
	Listening comprehension			0.06	7.53	<0.001

Note: GIA = General Intellectual Ability. In this paper GIA is conceptualized as a manifest indicator that represents statistical or *psychometric g*, and not a *theoretical* or *psychological g* latent brain-based biological or cognitive process dimension (see Fried 2020; McGrew et al. 2022).

Table 3. Standardized Beta Coefficients for Models 4, 5, and 6.

	Predictors					
Model	**Age**	***g***	**Oral Expression**	**Listening Comprehension**	**Test Directions**	**Test Directions × Age**
4	**0.029**	**0.656**	**0.047**		0.021	**−0.026**
5	**0.031**	**0.645**		**0.055**	0.021	**−0.026**
6	**0.029**	**0.637**	**0.030**	**0.039**	0.021	**−0.026**

Note: Bold values indicate significant predictors in the models.

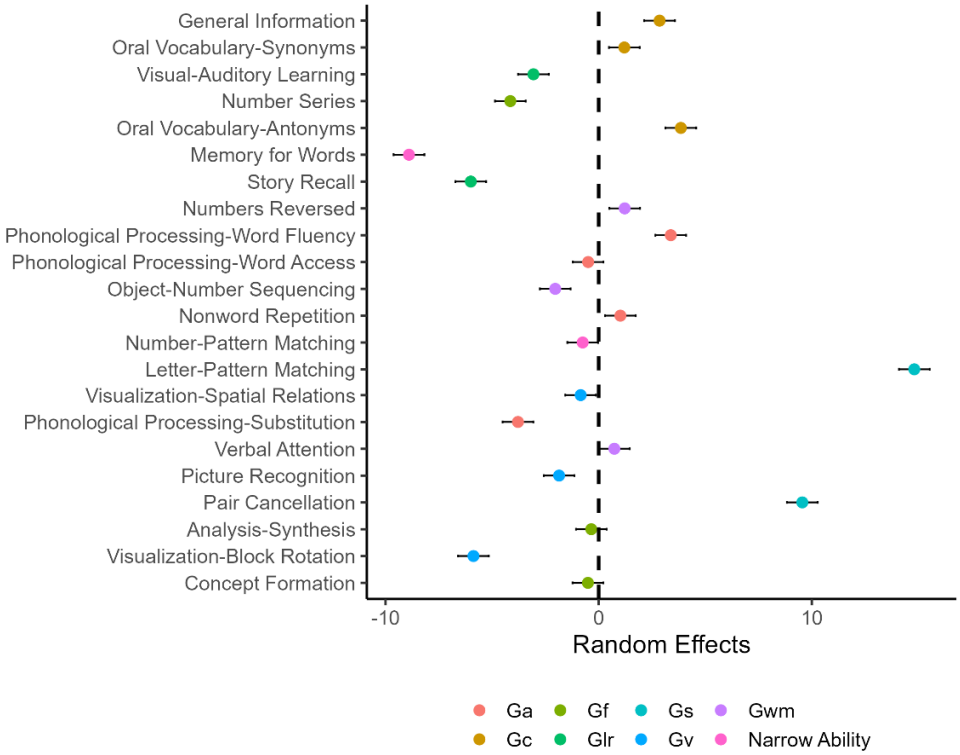

Figure 1. Estimated random effects and 95% confidence intervals for the tests in Model 1 (note: The dashed line represents the overall average score. The subtests on the *y*-axis are sorted in descending order by linguistic demand of the test directions).

When the variable *test directions* and its interaction with age were added to the model (Model 3), the variance estimate for level 2 (i.e., tests) remained relatively unchanged. This finding was mainly because *test directions* was not a significant predictor of the variation in the subtest scores. An interesting finding was that despite the variable *test directions* not being a significant predictor of cognitive test performance, its interaction with age was a statistically significant, negative predictor of cognitive test performance. This finding indicates that the impact of *test directions* was larger for younger students who are expected to have lower language proficiency compared with older students. This trend—*test directions* not being a significant predictor of individual cognitive test performance, but its interaction with age being a significant predictor of individual cognitive test performance—continued as the two additional variables were added to the model (see Models 4, 5, and 6 in Table 2). Models 4 and 5 showed that both *oral expression* and *listening comprehension* were statistically significant, positive predictors of individual cognitive test performance, after controlling for the effects of age and general intelligence.

Another important finding was that when both *oral expression* and *listening comprehension* were included in the model (Model 6), the two predictors remained statistically significant. The standardized beta coefficients produced from models 4 and 5 suggest that oral expression and listening comprehension were stronger predictors of cognitive test

performance than students' age, but relatively weaker predictors compared with the GIA score (see Table 3). When these four predictors (age, GIA, oral expression, and listening comprehension) were used together in the final model (Model 6), the GIA remains the strongest predictor, followed by listening comprehension (receptive language ability), oral expression (expressive language ability), and age, based on the standardized coefficients. The standardized beta coefficients also provided additional information regarding the negligible contribution of test directions within the models.

4. Discussion

The results of this study represent an integration of multiple pieces of empirical research completed over several years and across numerous studies. When examined in isolation, it appeared that examinee characteristics and test characteristics (e.g., test directions) *both* played meaningful roles in the administration and interpretation of cognitive tests for students. However, the integration of these potential influences into a single model has led to a surprising finding: the influence of examinee characteristics appears to eliminate the contribution of this test characteristic (i.e., the linguistic demand of test directions) on test performance. In addition, the influence of language ability, particularly receptive language ability, is more influential than age on cognitive test performance. This last point highlights the importance of considering language abilities when assessing students' cognitive abilities.

4.1. Examinee versus Test Characteristics

There appears to be increasing evidence that previous claims related to test characteristics (e.g., Cormier et al. 2011) no longer apply to contemporary tests of cognitive abilities. For example, the observed lack of a relationship between test directions and performance on the WJ IV in our study appears to draw a parallel with comments made by Cormier, Wang, and Kennedy, as they observed a "reduction in the relative verbosity of the test directions" when comparing the most recent version of the Wechsler Intelligence Scale for Children (WISC; Wechsler 2014) to the previous version (Wechsler 2003). These findings, in addition to the generation of clear guidelines for test development (e.g., AERA et al. 2014), both support the notion that large-scale, standardized measures include greater evidence of validity for the diverse population from which they are normed. This, in turn, likely contributes to increased fairness for the students that are assessed using these tests.

Despite the advances in test development, considerable challenges in assessing EL students remain for psychologists. One such challenge is assessing the cognitive abilities of the growing number of students who are considered ELs; limited English proficiency can lead to linguistically biased test results, which would lead to a misrepresentation of the examinee's true cognitive abilities. To eliminate this potential source of bias, psychologists testing EL students could consider examinee characteristics *before* administering a standardized measure of cognitive ability. This idea is not new. More than a decade ago, Flanagan et al. (2007) noted the critical need for psychologists to collect information regarding students' level of English proficiency, and the level of English required for the student to be able to comprehend test directions, formulate and communicate responses, or otherwise use their English language abilities within the testing process. Nonetheless, the results of our study provide an *empirical basis* in support of this broad recommendation.

4.2. Assessing English Language Abilities

The primary reason for assessing an examinee's English language skills is to determine if the examinee has receptive and expressive language skills that are comparable to the measure's normative sample. However, relying on one's clinical judgment when assessing an examinee's expressive and receptive language abilities is not likely to lead to positive outcomes. If practitioners only rely on their own judgment to determine the examinee's receptive and expressive language abilities, this could lead to either an under- or over-estimation of these abilities. An under-estimation could occur if the examiner deviates

from the standardized administration because they do not believe that the examinee has understood the directions. Thus, the linguistic demand of the actual, standardized test directions is potentially reduced. An over-estimation may occur if the examiner disregards the influence of the examinee's language abilities during testing and the results are interpreted the same for all examinees, regardless of their language abilities. In either case, an examiner who relies on their own judgment introduces unnecessary error into the assessment process. Therefore, especially in the context of testing EL students, practitioners should collect *data* on the receptive and expressive language abilities of examinees, so they can more accurately and reliably consider the potential influence of these variables on test performance.

Testing both expressive and receptive language abilities is critical for several other reasons. First, the results of the current study suggest that both make unique contributions to cognitive test performance. Second, a student's receptive and expressive language abilities are not always at the same level of proficiency. Moreover, although a student's conversational level of English language proficiency could be perceived to be relatively consistent with their peers', their level of academic language proficiency may not be sufficient to fully benefit from classroom instruction or understand test directions to the same extent of a native English language speaker (Cummins 2008).

Some practitioners may have concerns regarding the additional testing time required to administer, score, and interpret performance on language ability tests. Flanagan et al. (2013) addressed this concern well, as they explained:

> Irrespective of whether test scores ultimately prove to have utility or not, practitioners must endeavor to ascertain the extent to which the validity of any obtained test scores may have been compromised prior to and before any interpretation is offered or any meaning assigned to them. (p. 309)

Therefore, not only would this process be consistent with the aforementioned standards, but it would also lead to recommendations that are better informed and tailored to individual examinee characteristics.

4.3. Limitations and Future Research

This study was the first to integrate multiple sources of influence on test performance using a large, representative sample of the United States school-age population. However, the study was not without limitations, some of which may inform future efforts to continue with this line of inquiry. First, although the large, representative sample used in this study contained a wide range of language ability levels, it did not contain many EL students. Continuing to investigate the extent to which examinee characteristics matter, particularly within a well-defined sub-population of EL students, would better inform best practices with respect to the assessment of EL students.

The influence of examiner characteristics was noted as one of the three potential contributors to test performance. However, only examinee characteristics and test characteristics were included in the models produced for this study. Although examiner characteristics likely precede the psychoeducational assessment process with respect to assessing one's ability to engage in culturally sensitive practices and mastering the various aspects of test administration, scoring, and interpretation, their potential influence on test performance for EL students could be the focus of future research.

Author Contributions: Conceptualization, D.C.C. and O.B.; methodology, O.B.; formal analysis, O.B.; data curation, K.S.M.; writing—original draft preparation, D.C.C., O.B., K.S.M. and K.K.; writing—review and editing, D.C.C., O.B. and K.S.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: Ethical review and approval were waived for this study, due to an archival dataset being used for the analyses.

Informed Consent Statement: Participant consent was waived due to the researchers only receiving a de-identified dataset for their secondary analysis.

Data Availability Statement: The data presented in this study are not publicly available because they are confidential and proprietary (i.e., owned by the WJ IV publisher). Requests to access the data should be directed to Riverside Insights.

Conflicts of Interest: The authors declare no conflict of interest, with one exception: K.S.M. is a co-author of the WJ IV battery and discloses that he has a financial interest in the WJ IV.

References

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). 2014. *Standards for Educational and Psychological Testing*, 2nd ed. Washington, DC: American Psychological Association.

Barbey, Aron Keith. 2018. Network Neuroscience Theory of Human Intelligence. *Trends in Cognitive Sciences* 22: 8–20. [CrossRef] [PubMed]

Bates, Douglas, Martin Maechler, Benjamin M. Bolker, and Steven C. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48. [CrossRef]

Brown, Robert T., Cecil R. Reynolds, and Jean S. Whitaker. 1999. Bias in mental testing since bias in mental testing. *School Psychology Quarterly* 14: 208–38. [CrossRef]

Cormier, Damien Clement, Kevin S. McGrew, and Jeffrey J. Evans. 2011. Quantifying the "degree of linguistic demand" in spoken intelligence test directions. *Journal of Psychoeducational Assessment* 29: 515–33. [CrossRef]

Cormier, Damien Clement, Kevin S. McGrew, and James E. Ysseldyke. 2014. The influences of linguistic demand and cultural loading on cognitive test scores. *Journal of Psychoeducational Assessment* 32: 610–23. [CrossRef]

Cormier, Damien Clement, Kun Wang, and Kathleen E. Kennedy. 2016. Linguistic Demands of the Oral Directions for Administering the WISC-IV and WISC-V. *Canadian Journal of School Psychology* 31: 290–304. [CrossRef]

Cormier, Damien Clement, Okan Bulut, Deepak Singh, Kathleen E. Kennedy, Kun Wang, Alethea Heudes, and Adam J. Lekwa. 2018. A Systematic Examination of the Linguistic Demand of Cognitive Test Directions Administered to School-Age Populations. *Journal of Psychoeducational Assessment* 36: 337–53. [CrossRef]

Cummins, Jim. 2008. BICS and CALP: Empirical and theoretical status of the distinction. In *Encyclopedia of Language and Education*, 2nd ed. Edited by Brian Street and Nancy H. Hornberge. Volume 2: Literacy. New York: Springer, pp. 71–83.

Detterman, Douglas K., Elizabeth Petersen, and Meredith C. Frey. 2016. Process overlap and system theory: A simulation of, comment on, and integration of Kovacs and Conway. *Psychological Inquiry* 27: 200–4. [CrossRef]

Flanagan, Dawn P., Samuel O. Ortiz, and Vincent C. Alfonso. 2007. *Essentials of Cross-Battery Assessment with C/D ROM*, 2nd ed. New York: Wiley.

Flanagan, Dawn P., Samuel O. Ortiz, and Vincent. C. Alfonso. 2013. *Essentials of Cross-Battery Assessment*, 3rd ed. New York: John Wiley & Sons.

Fried, Eiko I. 2020. Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry* 31: 271–88. [CrossRef]

Harry, Beth, and Janette Klingner. 2014. *Why Are So Many Minority Students in Special Education?* New York: Teachers College Press.

Helms, Janet E. 1992. Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist* 47: 1083–101. [CrossRef]

Jensen, Arthur R. 1980. *Bias in Mental Testing*. New York: Free Press.

Kovacs, Kristof, and Andrew R. A. Conway. 2019. What is IQ? Life beyond "general intelligence". *Current Directions in Psychological Science* 28: 189–94. [CrossRef]

Kranzler, John H., Cindy G. Flores, and Maria Coady. 2010. Examination of the cross-battery approach for the cognitive assessment of children and youth from diverse linguistic and cultural backgrounds. *School Psychology Review* 39: 431–46. [CrossRef]

McGrew, Kevin S., Erica M. LaForte, and Fredrick A. Schrank. 2014. *Technical Manual: Woodcock-Johnson IV*. Rolling Meadows: Riverside Publishing.

McGrew, Kevin, Scott L. Decker, W. Joel Schneider, and Okan Bulut. 2022. IQ test structural research—To g or not to g? Psychometric network analysis of CHC measures of intelligence. *Unpublished manuscript*.

Ortiz, Samuel O. 2019. On the Measurement of Cognitive Abilities in English Learners. *Contemporary School Psychology* 23: 68–86. [CrossRef]

Ortiz, Samuel O., Salvador Hector Ochoa, and Agnieszka M. Dynda. 2012. Testing with culturally and linguistically diverse populations. In *Contemporary Intellectual Assessment: Theories, Tests and Issues*. Edited by Dawn P. Flanagan and Patti L. Harrison. New York: Guilford Press, pp. 526–52.

Protzko, John, and Roberto Colom. 2021. Testing the structure of human cognitive ability using evidence obtained from the impact of brain lesions over abilities. *Intelligence* 89: 101581. [CrossRef]

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Computer Software. Vienna: R Foundation for Statistical Computing.

Reynolds, Cecil R. 2000. Why is psychometric research on bias in mental testing so often ignored? *Psychology, Public Policy, and Law* 6: 144–50. [CrossRef]

Reynolds, Matthew R., and Christopher R. Niileksela. 2015. Test Review: Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock-Johnson IV Tests of Cognitive Abilities*. *Journal of Psychoeducational Assessment* 33: 381–90. [CrossRef]

Schrank, Fredrick A., Kevin S. McGrew, and Nancy Mather. 2014a. *Woodcock-Johnson IV Tests of Cognitive Abilities*. Rolling Meadows: Riverside Publishing.

Schrank, Fredrick A., Kevin S. McGrew, and Nancy Mather. 2014b. *Woodcock-Johnson IV*. Rolling Meadows: Riverside Publishing.

Schrank, Fredrick A., Kevin S. McGrew, and Nancy Mather. 2014c. *Woodcock-Johnson IV Tests of Achievement*. Rolling Meadows: Riverside Publishing.

Schrank, Fredrick A., Kevin S. McGrew, and Nancy Mather. 2014d. *Woodcock-Johnson IV Tests of Oral Language*. Rolling Meadows: Riverside Publishing.

Skiba, Russell J., Ada B. Simmons, Shana Ritter, Ashley C. Gibb, M. Karega Rausch, Jason Cuadrado, and Choong-Geun Chung. 2008. Achieving equity in special education: History, status, and current challenges. *Exceptional Children* 74: 264–88. [CrossRef]

Styck, Kara M., and Marley. W. Watkins. 2013. Diagnostic utility of the Culture-Language Interpretive Matrix for the Wechsler Intelligence Scales for Children—Fourth Edition among referred students. *School Psychology Review* 42: 367–82. [CrossRef]

Tychanska, Joanna. 2009. Evaluation of Speech and Language Impairment Using the Culture-Language Test Classifications and Interpretive Matrix. Doctoral dissertation, Saint John's University, New York, NY, USA. *Unpublished*.

van der Maas, Han L. J., Alexander O. Savi, Abe Hofman, Kees-Jan Kan, and Maarten Marsman. 2019. The network approach to general intelligence. In *General and Specific Mental Abilities*. Edited by Denis J. McFarland. Newcastle: Cambridge Scholars Publishing, pp. 108–31.

Wasserman, John D. 2019. Deconstructing CHC. *Applied Measurement in Education* 32: 249–68. [CrossRef]

Wechsler, David. 2003. *Wechsler Intelligence Scale for Children*, 4th ed. San Antonio: The Psychological Corporation.

Wechsler, David. 2014. *Wechsler Intelligence Scale for Children*, 5th ed. Bloomington: Pearson Education Inc.

Woodcock, Richard W., Kevin S. McGrew, and Nancy Mather. 2001. *Woodcock-Johnson III*. Itasca: Riverside Publishing.

Journal of
Intelligence

MDPI

Article

Measurement Efficiency of a Teacher Rating Scale to Screen for Students at Risk for Social, Emotional, and Behavioral Problems

Gino Casale [1,*], Moritz Herzog [1] and Robert J. Volpe [2]

[1] School of Education, Institute for Educational Research, University of Wuppertal, Gaußstraße 20, 42119 Wuppertal, Germany
[2] Department of Applied Psychology, Northeastern University, Boston, MA 02115, USA
* Correspondence: gcasale@uni-wuppertal.de

Abstract: Teacher rating scales are broadly used for psycho-educational assessment in schools. In particular, they play an important role in screening students for social, emotional, and behavioral problems. In order to optimize the efficiency of these measures, it is important to minimize the number of items comprising them while maintaining sound psychometric characteristics. This study examines the measurement efficiency of a teacher rating scale for student social, emotional, and behavioral risk. The goal was to shorten an existing behavior screening tool. A total of 139 classroom teachers and 2566 students from Grades 1–6 (M_{age} = 8.96 years, SD = 1.61) participated in the study. In sum, 35 items assessing internalizing and externalizing behavior problems were analyzed applying the item response theory (generalized partial credit model). The results show that social, emotional, and behavioral risks can be captured with a total of 12 items. This reduction of almost 66% of the initial item pool would take teachers about 90 s to fill out for one student. Thus, the rating scale can be used by teachers in an efficient yet psychometrically sound manner.

Keywords: universal screening; item response theory; behavior problems; school-based assessment

Citation: Casale, Gino, Moritz Herzog, and Robert J. Volpe. 2023. Measurement Efficiency of a Teacher Rating Scale to Screen for Students at Risk for Social, Emotional, and Behavioral Problems. *Journal of Intelligence* 11: 57. https://doi.org/10.3390/jintelligence11030057

Received: 5 November 2022
Revised: 6 March 2023
Accepted: 16 March 2023
Published: 19 March 2023

1. Introduction

1.1. Social, Emotional, and Behavioral Competencies in Children and Adolescents

The social, emotional, and behavioral development of children and adolescents plays a central role in primary education. Social and emotional competence is a broad and multidimensional construct for which many different operationalizations and models exist (Berg et al. 2019). At its core, social–emotional competence refers to interpersonal and intrapersonal skills in the emotional (e.g., emotion knowledge and emotion regulation), social (e.g., social problem solving, processing social cues), and cognitive (e.g., executive functions) domains (Berg et al. 2019).

Among other things, these skills are associated with academic performance, school success, and the development of psychosocial disorders (Aviles et al. 2006; Domitrovich et al. 2017). A large proportion of all school-age children and adolescents shows significant impairments in social, emotional, and behavioral development. Depending on the definition used and the informants involved in generating estimates, approximately 12–18% of children and adolescents with emotional and behavioral disorders can be identified internationally (Kovess-Masfety et al. 2016; Polanczyk et al. 2015). Among these, internalizing disorders, such as anxiety, occur more frequently than externalizing difficulties, such as attention-deficit/hyperactivity disorder (ADHD; Kovess-Masfety et al. 2016).

A variety of school-based interventions can promote social, emotional, and behavioral competencies in students. In three meta-analyses (Durlak et al. 2011; Korpershoek et al. 2016; Sklad et al. 2012), building prosocial behavior, reducing behavior problems, and increasing academic achievement was effective with a small effect size; increasing social–emotional skills was effective with a medium effect size.

However, there is often a significant gap between the initial presence of a student's social, emotional, and behavior problems and the provision of school-based interventions (Daniels et al. 2014). It is estimated that only about 20–30% of all children and adolescents with problems in social, emotional and behavioral development receive systematic support in terms of prevention or intervention (Langer et al. 2015). Although these numbers differ between countries, this "service gap" (Forness et al. 2012, p. 3) is widespread and concerning. One reason for this "underservice" is that many students with problems in their social, emotional, and behavioral development remain unidentified and their problems are not recognized until they already correspond to symptoms of a clinical disorder (e.g., Breitenstein et al. 2009). This problem precludes the application of early support services that have been shown to be effective in preventing the escalation of developmental trajectories (e.g., Durlak et al. 2011). Alternatively, early identification of the aforementioned problems can promote prevention and counteract the development of mental disorders (e.g., Costello 2016).

1.2. Early Identification of Social, Emotional, and Behavioral Risk in Students

Both the externalizing and internalizing behaviors of students are significant indicators of the social and emotional competence of children and adolescents. Externalizing behavior problems have a significant impact on positive social interactions in the classroom and disrupt learning and teaching processes (Lane et al. 2014). Therefore, these behavioral problems are often better and more accurately identified by teachers than internalizing behavioral problems, which are often overlooked (e.g., Dwyer et al. 2006; Hartman et al. 2017). For this reason, among others, it is important to provide teachers with tools that can be used for the early identification of students' externalizing and internalizing behavioral problems (Splett et al. 2019).

Many different approaches exist for the assessment of social, emotional, and behavioral characteristics in children and adolescents, e.g., behavioral observations, test batteries, or more innovative approaches such as situational judgement tests or forced choice assessments (Halle and Darling-Churchill 2016). These methods usually show acceptable to good psychometric characteristics, but are often very time-consuming in regard with preparation, implementation, and evaluation, which is incompatible with everyday school routines. As such, they may not be suitable for the universal screening of at-risk students.

For an initial assessment of whether students are exhibiting problems in the social, emotional, and behavioral domains, universal screening methods for student behavioral problems have proven effective within a decision-making process in an evidence-based assessment (Volpe and Briesch 2018). Universal behavior screening tools "are conducted with all students in a classroom [...] to identify those at-risk of behavioral difficulties or emotional and behavioral disorders (EBD) who could potentially benefit from specific instruction or intervention" (Glover and Albers 2007, p. 118). Eklund et al. (2009) showed that the use of universal screening procedures identified more than twice as many at-risk students as other psychoeducational assessment practices. Ideally, a consequent result of this early detection of at-risk students is the provision of interventions at the first sign of these problems (Volpe et al. 2010).

In general, universal behavioral screenings work by having teachers complete ratings for each student. The results can be used to make decisions regarding student risk for developing severe social–emotional behavioral problems. However, several studies show that far fewer than half of all schools and teachers systematically screen their students for social, emotional, and behavioral risks (Bruhn et al. 2014; Dineen et al. 2022; Glover and Albers 2007; Wood and Ellis 2022). This still strongly underutilized use of universal screenings can be attributed in part to the overly broad scope of many standardized screening instruments, which tend to discourage teachers from using them (Burns and Rapee 2019; Volpe et al. 2018). One important predictor of the implementation of universal screening procedures is the teachers' attitudes towards screening (Moore et al. 2022). Teachers' attitudes towards universal screening are mainly affected by the required resources for implementation, espe-

cially the time teachers need for completion (Briesch et al. 2017; Kauffman 1999). Therefore, one critical feature of universal screening tools should be that they are highly time-efficient, but still beneficial for practical use in schools.

An established procedure for the time-efficient screening of social, emotional, and behavioral risks in children in school is multiple-gating (Walker et al. 2014). The basic idea behind multiple-gating procedures is to progressively narrow down the pool of potential at-risk students by using increasingly rigorous methods at each successive gate. This approach is also promoted as best practice in screening in school contexts (Whitcomb and Merrell 2013), and has been shown to be superior to using a procedure involving a single measure (Kilgus et al. 2018). Efficiency is gained in this approach if time-efficient measures are used in earlier gates to rule out typically developing students with more time-intensive methods used for the remaining students. Multiple-gating procedures often have three stages (see Stiffler and Dever 2015): first, the teacher nominates students who the teacher subjectively perceives as exhibiting social, emotional, and behavioral problems. A comparatively short broadband rating scale is then completed for the students who advance to the second gate. A third gate could either consist of a systematic direct observation of a small pool of students or a more comprehensive rating scale.

1.3. Measurement Efficiency of Universal Behavior Screenings

Following Glover and Albers (2007) and Volpe and Briesch (2018), universal behavior screening procedures should meet three essential requirements: (1) Appropriateness for the intended use (i.e., alignment with constructs of interests and theoretical and empirical support); (2) Technical adequacy of the tool (i.e., psychometric properties); and (3) Usability of the tool (i.e., cost–benefit ratio, acceptability, and utility of outcomes). With regard to school-based universal screening, the appropriateness for the intended use is given if the tool provides timely and useful information regarding the levels of risk for all students (Daniels et al. 2014). In the school context, the constructs of interest are not clinically relevant symptom scales, but rather behavioral scales that capture problems in social, emotional, and behavioral dimensions (see Volpe et al. 2018). Technical adequacy indicates that the screener demonstrates acceptable reliability, validity, and accuracy in the early identification of at-risk children (i.e., classification accuracy). Usability implies that: (a) The tool is feasible and acceptable to stakeholders; and (b) The results of the screener guide the selection of interventions (Glover and Albers 2007).

This third category of usability also includes the aspect of measurement efficiency (e.g., Anthony et al. 2016). By measurement efficiency we mean that the preparation, implementation, and interpretation of the measurement instrument are carried out with the least possible time effort while obtaining the best possible psychometric information (Anthony et al. 2016). With reference to behavior rating scales, this means that the number of items to be completed is minimized, but these items are still representative for the underlying latent constructs, and thus, the results can be used meaningfully to identify at-risk students (Glover and Albers 2007). If these psychometric requirements are met, the results of the screening can be used to distinguish between students with and without social, emotional, and behavioral risk.

In order to make the best selection of items for these purposes from a test theory perspective, it is important to obtain the most comprehensive and accurate information possible. Item response theory (IRT; e.g., Wilson 2004) is suitable for this purpose. In the context of IRT, the difficulty of the items (as manifest variables) is examined in relation to the actual trait expression of the subjects (as latent variables). For universal screening, this means the social, emotional, and behavioral problems of a student (latent trait) and the specific items (manifest traits) correspond accordingly (Anthony et al. 2016). IRT analyses could be used to map how well the items differentiate between different levels of competence (in this case, between students with and without risk). This approach also allows an analysis of which items are particularly salient and meaningful in classifying

between at-risk and non-at-risk students, so that the results can be used for optimal item selection and reduction (Hambleton 2000).

1.4. The Current Study

The current study represents a re-analysis of data published by Volpe et al. (2020) with results from using the integrated teacher reporting form (ITRF; Volpe and Fabiano 2013) to improve measurement efficiency for social, emotional, and behavioral risk. The instrument is considered a well-established universal screening for primary school students that includes 35 items related to internalizing and externalizing classroom behaviors, such as depressive behavior (AD), socially withdrawn behavior (SW), oppositional/disruptive behavior (OPD) and academic productivity behavioral problems (APP). The aim of the present study is to increase the measurement efficiency of the scale by reducing the number of items to a minimum level required to accurately discriminate between at-risk and non-at-risk students. More specifically, we were interested in retaining the items of the full ITRF that:

(a) Discriminate best between children with low and high levels of behavioral problems; and

(b) Are sensitive to students with above-average behavioral problems, but not necessarily very high problems. As students with very high levels of behavioral problems are the most likely to be identified by teachers (even without an assessment tool), early universal screening should detect even mild-to-moderate behavioral problems (Kendziora 2004).

While meeting the above-mentioned criteria, we seek to delineate a shortened version of the ITRF, which is comparable to the full-length version in regard to its ability to discriminate students with and without significant behavioral problems.

2. Materials and Methods

2.1. Participants and Setting

A total of 10 inclusive primary schools, 2 inclusive secondary schools, and 3 special schools from one school district in the federal state North Rhine Westphalia (NRW; Western Germany) participated in the study. In sum, 139 classroom teachers completed the questionnaires for 2566 students (48.2% female). The mean age of the teachers was 43.00 years (SD = 9.28), with a mean teaching experience of 15.84 years (SD = 8.96). The mean age of the student sample was 8.96 years (SD = 1.61), with a range from 6 to 15 years. The majority of the students was from Grades 1 to 4 (91.2%), 8.8% were from Grades 5 and 6. Regarding gender, 90.4% of the teachers were female. Information about the study and the data collection processes were provided by a member of the research team at a school principal meeting and additional personal communication (e.g., phone calls and mailing) before the data collection started. All schools received a packet containing ITRF forms, and an additional form to record the sociodemographic characteristics of students. Each individual classroom teacher completed both forms for all the students in the classroom and sent them back to the investigators.

2.2. Instrument—The Integrated Teacher Report Form (ITRF)

The ITRF was initially developed to assess the externalizing behavioral problems of primary school students in the classroom (Volpe and Fabiano 2013). The English-language ITRF was translated into German and adapted and validated for use in both a long and a short version (Casale et al. 2018; Volpe et al. 2018). In addition, the instrument was expanded and validated with items referring to internalizing classroom behaviors (Volpe et al. 2020). This version assesses student externalizing and internalizing classroom behaviors that indicate a social, emotional, and behavioral risk (Volpe et al. 2020). It consists of 35 items (see Appendix A) measuring academic productivity problems (8 items), oppositional/disruptive behavior (8 items), anxious/depressive behavior (11 items), and social withdrawal (8 items). The ITRF is part of the Integrated Screening and Intervention System (Volpe and Fabiano 2013), which incorporates universal screening, intervention, and

behavioral progress monitoring. Numerous studies support its factorial validity, internal consistency, retest reliability (Daniels et al. 2014; Volpe et al. 2018, 2020), construct validity (Casale et al. 2019), and cross-cultural equivalence (Casale et al. 2018). In particular, those studies examined how the ITRF relates to other established behavioral screening measures. However, those studies only included the externalizing scales of the ITRF. Daniels et al. (2014) tested convergent and discriminant validity and used a symptom-based behavioral assessment for teachers in addition to the ITRF (brief problem monitor; Achenbach et al. 2011). High correlations between content-like constructs and low correlations between content-distant constructs underscore the construct validity of the ITRF. For the German-language version, the classification accuracy and predictive validity for identifying a problem of the ITRF was analyzed (Volpe et al. 2018). For this purpose, the Teacher Report Form of the Child Behavior Checklist (TRF-CBCL; Achenbach et al. 2008) was used as the criterion measure. The calculation of receiver operating curves (ROC) and positive as well as negative predictive values (PPV & NPV) indicated a high diagnostic accuracy for all scales of the externalizing ITRF (AUC .85–.94). For all scales, NPVs were substantially higher than PPVs, which is acceptable for a screening procedure because more students are selected for intervention than are actually prevalent psychosocial problems (Volpe et al. 2018). Finally, in another study with the German-language ITRF, convergent and discriminant validity were analyzed using a multitrait–multimethod correlation matrix and a correlated trait–correlated method minus 1 model to separately analyze the influence of the constructs (learning-related/attentive behavioral problems, oppositional/disruptive behavioral problems) and the methods (ITRF, additional assessment procedure) on the resulting scores (Casale et al. 2019). The additional screenings were the strengths and difficulties questionnaire (SDQ; Goodman 1997), the TRF-CBCL, and the *Lehrereinschätzliste für Sozial- und Lernverhalten* (LSL; Petermann and Petermann 2013; teacher assessment schedule for social and learning behavior). The results demonstrate that the theoretically postulated correlations can be mapped to the empirical data, in line with expectations, indicating convergent and discriminant validity. The variance of the ITRF values can be explained to a greater extent by the construct being measured than by method-specific influences, which also supports the construct validity of the ITRF. In addition, Volpe et al. (2018) conducted a systematic comparison of the externalizing ITRF with established German-language screening procedures (SDQ, TRF-CBCL, LSL) in terms of their usability for school-based use. The results demonstrate that except for the ITRF, none of the instruments are fully suitable for use in schools because they are either too symptom-orientated (TRF-CBCL), too comprehensive (TRF-CBCL, LSL), or not systematically linked to school-based interventions (SDQ, TRF-CBCL) (Volpe et al. 2018).

In this study, the participating classroom teachers completed the full-length ITRF for all the students in their classroom in order to precisely identify the problematic classroom behaviors raising most of the concern for the students. The teachers completed the ITRF items on a 4-point Likert scale (0 = behavior is not of concern, 1 = behavior is of slight concern, 2 = behavior is of moderate concern, and 3 = behavior is of strong concern).

2.3. Analysis Design

To identify the items of the full ITRF that discriminate well between students with low and with high levels of behavioral problems, and that measure especially slightly above the population mean, we applied item response theory (IRT) models, in particular the generalized partial credit model (GPCM). IRT models measure a latent trait (e.g., behavioral problems) on the same scale as the corresponding items (the theta (θ) continuum). That means that for each item, a location on the theta continuum can be estimated (Parameter β). In terms of questionnaires, this parameter can be interpreted as the likelihood with which raters will rate a higher score at this item (or "agreeability"). Given that IRT models are probabilistic models, the location on the theta continuum is defined as the level of the underlying trait at which the probability of being scored higher increases the most ($P(\theta)$). Given a limited amount of answer options (e.g., on a Likert-type scale), when items are

dichotomous (e.g., yes or no), IRT models only report one parameter of "agreeability"; however, when items are polytomous (e.g., never, sometimes, often, and very often), there are several thresholds estimated that indicate the level of the underlying trait at which the most probable answer changes (e.g., from never to sometimes). As these parameters (τ_i) indicate the borders between the most probable answers, there is one parameter less than for the answer options. The GPCM has the advantage that the steepness in which the probability of being scored higher increases can be differentiated between the items (Parameter α) (Muraki 1997). This parameter indicates how strongly the item discriminates between persons with a high trait and a low trait. The probability of multiple answers (e.g., in a Likert scale) across the theta range can be illustrated in the item characteristic curve (ICC). While dichotomous items only have on curve (e.g., for the category "right"), polytomous items have several curves—one for each answer option. Figure 1 shows a typical ICC for an item with four answer options and also illustrates the item parameters α and τ_i.

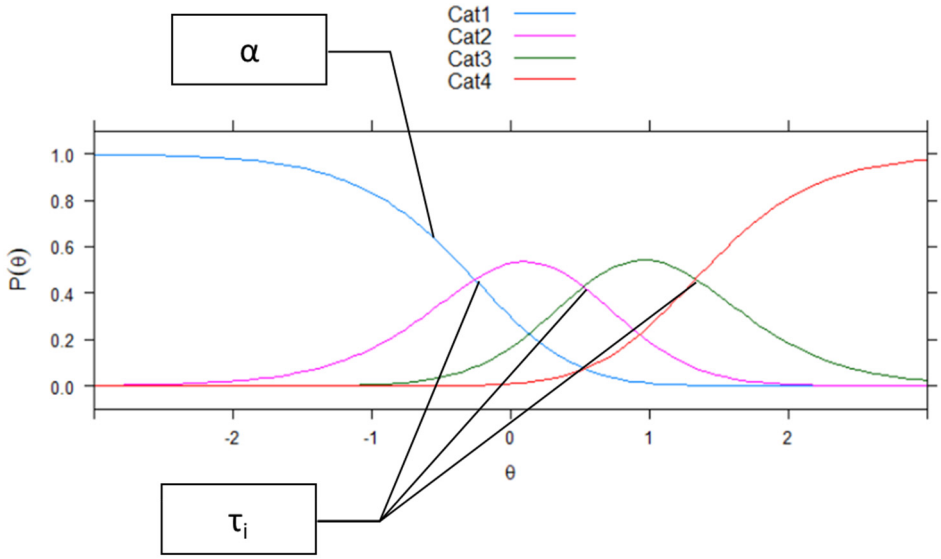

Figure 1. Typical ICC of a polytomous item and item parameters α (discrimination) and τ_i (threshold location). θ refers to the latent trait. P(θ) refers to the probability of the answer categories. The different colors of the curves refer to the different answer categories.

The IRT analyses are structured in two sections. First, the items of the full version of the ITRF were reduced. Based on the parameters of the GPCM, items showing the highest discrimination (values of α) and a comparably low "agreeability" (values of β and τ_i) were selected for retention. Since IRT models require that the items under investigation measure a unidimensional construct, items of the ITRF were divided into four subscales (AD, SW, OPD, and APP), as indicated by Volpe et al. (2020). The selected items for each subscale were taken as potential shortened versions of the full-length ITRF subscales. Second, the internal and external validity of the new versions was investigated. Internal validity was checked by Crombach's α. To investigate to what extent the full version and the shortened versions of the ITRF correspond, correlations between the sum scores were calculated.

All analyses were conducted in *R* (R Core Team 2022) using the packages *TAM* (Test Analysis Modules; Robitzsch et al. 2022) and *psych* (Procedures for Psychological, Psychometric, and Personality Research; Revelle 2022).

3. Results

In total, four GPCMs were employed, one for each subscale of the ITRF. Two main assumptions have to be fulfilled before applying IRT models to the data. First, the data has to be unidimensional. This means that the items included in the model cover the same construct. Usually, unidimensionality is investigated via factor analysis. Given the factor

analysis provided by Volpe et al. (2020), the four subscales of the ITRF are unidimensional and distinct from each other.

Second, the data have to be locally independent. That means that there are rarely covariations among the items. Typically, Q_3 statistics between the item pairs of a data set are used to check for local dependency (LD). There are different critical values of the Q_3 statistic discussed in the literature. However, 0.2 and 0.3 appear to be often used as critical values for LD (Christensen et al. 2017). To test for LD, item pairs were formed within the subscales of the ITRF. Of a total of 139 item pairs, 103 (74%) showed a Q3 statistic below 0.2, 28 item pairs (20%) had a moderate Q_3 between 0.2 and 0.3, and eight item pairs (6%) had a considerable LD with a Q_3 statistic above 0.3.

LD is a common problem in data that were rated by several individuals (Anthony et al. 2016; Wu 2017). LD in such cases is often caused by general tendencies (e.g., trend to the middle) and individual tendencies (e.g., leniency) in rating behavior (Wu 2017). Song (2019) showed that LD compromises the results of a GPCM only to a small degree. As the aim of this study was not to assess individuals' traits in detail, but to compare item characteristics, GPCMs still appear adequate.

The main basis for the item reduction in the four subscales of the ITRF was the degree of discrimination (α) and the item location (i.e., the range of the underlying trait where the item measures best; β). Based on the item characteristics, three items from each subscale were selected for the shortened version of the ITRF. Three selection criteria were applied: First, high discrimination between persons with low and high behavioral problems (high parameter α). Second, low item location within the latent trait continuum (low parameter β). Additionally, third, a small theta range in which "no difficulties" was the most probable answer category (low parameter τ_1). Table 1 comprises the information on discrimination, item location, and theta range for τ_1. Finally, in terms of content, we examined whether the items that met the aforementioned psychometric criteria also matched the underlying constructs in terms of content and were not too similar in content or redundant.

To check to what extent the shortened version of the ITRF is more sensitive in the middle theta range, test information curves (TICs) were plotted. TICs display the information an item (collection) provides across the theta range. The shape of a TIC can inform, in which theta range (e.g., little or severe behavioral problems) the focus of test information of an item collection lies. Figure 2 shows the TICs of the subscales and the full scale of the original ITRF and the shortened version. The TICs illustrate that the information focus of the shortened subscales AD, SW, and APP had shifted to the theta range of between 0 and 1 compared to the full versions. In the subscale OPD, the information focus had only slightly shifted to the theta range between 0 and 1. However, as in the subscale OPD, as the items that had the lowest localization on the theta range (parameter β) had already been selected, no further optimization would be possible. Regarding the full ITRF, the test information of the shortened version had slightly shifted to the theta range between 0 and 1.

To check if the shortened version of the ITRF has the same factor structure as the original version—and thus, if the subscales of the shortened version can be used to assess children's differential behavioral problems—a confirmatory factor analysis was employed (see Table 2). A model fit of the confirmatory factor analysis was acceptable (CFI = .954, TLI = .937, RMSEA = .075, $C.I._{RSMEA}$ = [.70–.80]) and all factor loadings were significant ($p < .001$). Factor loadings ranged from .65 to .86, and thus, confirmed the four-factor structure of the shortened version of the ITRF. The scale intercorrelations between both externalizing factors (r = .457, CI: .451–.464) and between both internalizing scales (r = .425, CI: .418–.432) were moderate (Table 3). The intercorrelations between the externalizing and internalizing factors were low to moderate (r = .160–.346). The internal consistency of the full scales and the subscales SW, OPD, and APP of the shortened version of the ITRF were good (Cronbach's α between .85 and .87). Additionally, the internal consistency of the subscale AD was acceptable (Cronbach's α = .73).

Table 1. Item parameters of the GPCMs for each subscale of the ITRF.

Item	α	β	τ_1	τ_2	τ_3
		Subscale AD			
I_2	1.665	1.669	−.648	.019	.629
I_7	1.745	1.522	−.692	.108	.584
I_8	1.665	1.623	−.553	.045	.509
I_9	1.774	1.468	−.773	.132	.641
I_10	1.282	1.859	−.080	−.082	.162
I_11	1.167	1.809	−.364	.117	.247
I_12	1.263	1.766	−.594	.077	.517
I_15	1.894	1.591	−.606	−.005	.611
I_17	1.345	1.789	−.613	.050	.562
I_19	1.848	1.833	−.333	.090	.243
I_23	2.357	1.549	−.647	.157	.490
		Subscale SW			
I_1	2.635	1.543	−.657	.064	.593
I_4	4.289	1.568	−.757	.096	.661
I_5	3.998	1.671	−.662	.162	.500
I_6	3.440	1.929	−.836	.094	.741
I_13	1.198	1.807	−.789	.063	.593
I_14	1.155	1.843	−.477	.112	.365
I_16	1.679	1.526	−.638	.019	.619
I_24	1.323	1.822	−.644	.189	.455
		Subscale OPD			
E_7	3.263	1.339	−.390	.043	.348
E_8	2.365	1.143	−.305	−.009	.314
E_9	2.136	.754	−.764	.098	.666
E_10	3.060	1.261	−.406	.026	.380
E_11	2.972	.823	−.715	.090	.625
E_12	1.879	.1292	−.491	−.035	.526
E_13	3.608	1.133	−.500	.059	.441
E_16	1.983	1.459	−.461	−.073	.534
		Subscale APP			
E_1	2.007	.632	−.642	.047	.595
E_2	2.203	.866	−.613	.067	.546
E_3	2.518	1.062	−.522	.023	.498
E_4	3.034	1.054	−.693	.061	.632
E_5	2.167	.654	−.844	.258	.685
E_6	1.849	1.198	−.477	−.029	.505
E_14	2.576	1.351	−.519	.045	.475
E_15	.920	1.174	−.649	.075	.573

Note. AD = anxious/depressed behavior; SW = social withdrawal; OPD = oppositional/defiant problems; APP = academic productivity problems; bold items were selected for the shortened version.

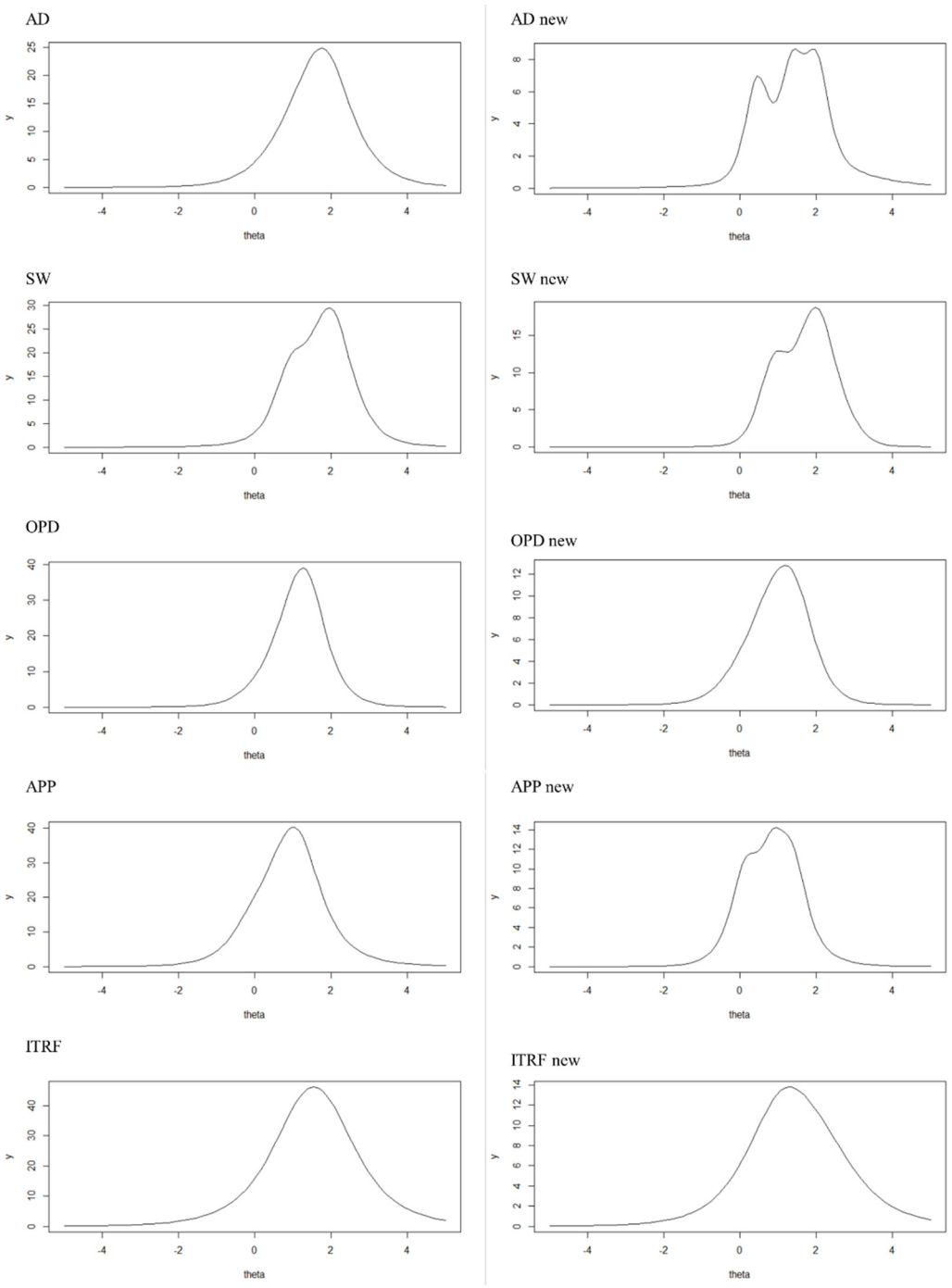

Figure 2. Test information curves for subscales and full questionnaire of the original (**left**) and shortened (**right**) version. Anxious/depressed behavior; SW = social withdrawal; OPD = oppositional/defiant problems; APP = academic productivity problems; ITRF = integrated teacher report form.

Table 2. Item factor loadings and reliability of the shortened ITRF.

Item	α	β
Anxious/Depressive	**.73**	
Appears unhappy or sad		.76
Complains or whines		.65
Spends a lot of time worrying		.65
Social Withdrawal	**.87**	
Avoids social interactions		.86
Prefers to play alone		.84
Does not respond to others' attempts to socialize		.80
Oppositional/Defiant Behavior	**.85**	
Disrupts others		.83
Has conflicts with peers		.81
Makes irrelevant comments		.80
Academic Productivity Problems	**.86**	
Does not complete classwork on time		.84
Does not start assignments independently		.91
Does not turn in class assignments		.74
Model Fit		
χ^2 = 739.748, df = 48, p = .000; CFI = .954, TLI = .937, RMSEA = .075, C.I.RSMEA = [.70–.80]		

Note. α = Cronbach's alpha; β = standardized factor loadings; df = degrees of freedom; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval.

Table 3. Factor correlations of the full and the short ITRF.

	Short ITRF				Full ITRF			
	AD	**SW**	**ODP**	**APP**	**AD**	**SW**	**ODP**	**APP**
Short ITRF								
AD		.56	.43	.41	.91	.59	.36	.34
SW			.18	.34	.47	.78	.16	.32
ODP				.52	.33	.27	.96	.50
APP					.34	.54	.42	.94
Full ITRF								
AD						.69	.37	.38
SW							.26	.52
ODP								.51

Note. AD = anxious/depressed behavior; SW = social withdrawal; OPD = oppositional/defiant problems; APP = academic productivity problems; all correlations were significant (p < .001).

In a final step, the concordance of the full scale and the subscales of the original and the shortened version of the ITRF were investigated. For all subscales and the full scale, the shortened and the original version correlated strongly (r > .78).

4. Discussion

The aim of this study was to maximize the measurement efficiency of a teacher rating scale for the school-based assessment of social, emotional, and behavioral risk in students.

IRT models were applied in order to analyze the potential to reduce the number of items of a well-established universal screening scale, the ITRF (Volpe and Fabiano 2013). The test information of the shortened version was supposed to be more focused on the theta range between 0 and 1 in order to be more sensitive to children with moderate social, emotional, or behavior problems. Finally, the shortened version had to measure similarly to the original version of the ITRF, including the factor structure. The shortened version proposed in this study meets all these criteria.

Our analyses indicate that the social, emotional, and behavioral risk of students can be assessed with 12 items only (three items per construct), which is a reduction of almost 66% of the original scale. Speaking in terms of time, and assuming a processing time of the original ITRF of about 5 min per student, the time required to complete the scale for a student can be reduced to about 90 s. For a universal screening of an entire school class of approximately 25 students, this means that the ITRF can be completed for all students in less than 40 min. It is thus ideally suited for a first time efficient yet psychometrically high-quality step in multiple-gating assessment. In a second gate, the longer ITRF could then be used for a more detailed clarification of the problems. Compared to the original ITRF, the teacher nomination step could, thus, be replaced by the systematic short screening developed here. Given this lower effort, the shortened version of the ITRF is more likely to be used in schools within multiple-gating procedures. Therefore, it contributes to the implementation of the regular assessment of children's individual social and emotional development, as well as their specific needs.

The current study showed that reducing items and shortening questionnaires is applicable without sacrificing psychometric rigor. Previous studies from different fields have given similar examples on how a questionnaire can be reduced (Anthony et al. 2016; Becker et al. 2007; Chiesi et al. 2018; Volpe et al. 2011; Volpe and Gadow 2010). Based on these experiences, researchers developing questionnaires might always consider test efficiency and—if possible—prepare a short version for screening purposes in general.

The present re-analysis is a further step in the development of a well-implementable, school-based behavioral screening. The items identified here for the short version need to be investigated in future studies with a different sample with regard to their factorial validity, their external evidence (especially convergent and divergent validity in comparison with other established scales), and their predictive power for the identification of actual behavioral problems. This seems particularly relevant in light of the fact that the extensive evidence on the construct validity of the longer ITRF has predominantly worked with the externalizing scales. A more in-depth analysis of the internalizing scales is yet to be conducted.

The results can be discussed against the background of teachers' tendency to detect externalizing problems more easily than internalizing problems (Dwyer et al. 2006; Hartman et al. 2017). The focus of the test information shifted to a lower theta range (referring to less severe behavioral problems) stronger for internalizing than externalizing problems. Thus, the full versions of the externalizing scales, especially the OPD, were already strongly focused on a lower theta range, whereas the full internalizing scales focused more on a higher theta range (referring to students with severe internalizing problems). Selecting the items most sensitive for slightly above-average behavioral problems within the theta range of 0 to 1 affected the internalizing scales stronger than the externalizing scales. Moreover, the mean beta parameters of the internalizing items were higher than of the externalizing items. Lower beta parameters in the externalizing scales indicate that these items are more likely to be scored higher by teachers even if the behavioral problems are less severe. Conversely, higher beta parameters in the internalizing scales indicate that students need to have more severe internalizing behavioral problems for teachers to score the corresponding items higher. Thus, the results corroborate findings stating that teachers can detect externalizing problems better than internalizing problems.

Limitations

The findings of the current study should be interpreted in the context of at least four limitations. First, the item reduction was merely based on the GPCMs and the parameters for discrimination and location on the theta range. This procedure pays little respect to the content of the items. For example, including an expert rating regarding the most relevant items of the original version of the ITRF would provide a broader empirical basis for the item selection.

Second, the revalidation of the shortened version did not examine external validity with other measures (e.g., other questionnaires assessing social, emotional, and behavioral problems). Investigating the external validity of the shortened version of the ITRF would improve the interpretability of the results.

Third, predictive validity was not investigated. As the shortened version of the ITRF is supposed to serve as a screener for social, emotional, and behavioral problems, its predicative validity is of great interest. Information on the accuracy with which the shortened version of the ITRF can predict social, emotional, and behavioral problems with different severity would increase the interpretability of the instrument. Moreover, this information might convince more teachers to implement an early assessment of risk for social, emotional, or behavioral problems.

Fourth, in our resulting models, items showed considerable local dependencies (LD). Even if this is a common problem in individual teacher ratings (Anthony et al. 2016; Wu 2017) and LD compromises the results of a GPCM only to a small degree (Song 2019), the results might be caused by specific rater effects, such as general tendencies or halo effects (Wu 2017). A potential solution might be psychometric evaluation approaches that allow to consider rater effects in behavior rating scales such as the many-facet Rasch model (see Anthony et al. 2022) or generalizability theory (e.g., Briesch et al. 2014). However, those approaches attempt quite strict a priori design specifications, which were not applied in the current study.

5. Conclusions

The results of the present study indicate that the assessment of students' social, emotional, and behavioral risk is possible even with only a few items in the teacher rating. The scale used here is thus very well suited for the time-efficient measurement of students' classroom behavior (90 s). This enables teachers to integrate behavioral diagnostics into their daily school routine and to identify students' needs at an early stage in order to implement appropriate support services and prevent the development of psychosocial disorders. With the shortened version of the ITRF, applying early assessment of social, emotional, and behavioral development is facilitated in schools.

Author Contributions: Conceptualization, G.C. and R.J.V.; Methodology, G.C., M.H. and R.J.V.; Formal analysis, M.H.; Investigation, G.C. and R.J.V.; Writing – original draft, G.C. and M.H.; Supervision, G.C. and R.J.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Open Access Publication Fund of the University of Wuppertal.

Institutional Review Board Statement: Ethical review and approval were not required in accordance with the local legislation and institutional requirements. Following the school law and the requirements of the ministry of education of the federal state North Rhine Westphalia (Schulgesetz für das Land Nordrhein-Westfalen), school administrators decided in co-ordination with their teachers about participation in this scientific study.

Informed Consent Statement: Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements. Verbal informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We thank Michael Grosche for his nuanced and thoughtful feedback on the final draft of this article.

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

Appendix A. Items of the ITRF (Bold Items Were Selected for the Shortened Version)

E-1. Does not complete classwork on time (APD 1)
E-2. Does not start assignments independently (APD 2)
E-3. Missing or incomplete homework (APD 3)
E-4. Does not turn in class assignments (APD 4)
E-5. Does not correct own work (APD 5)
E-6. Fails to pack needed materials for home (APD 6)
E-7. Argues with teacher (OPP 1)
E-8. Loses temper (OPP 2)
E-9. Disrupts others (OPP 3)
E-10. Uses inappropriate language (OPP 4)
E-11. Has conflicts with peers (OPP 5)
E-12. Bossy (OPP 6)
E-13. Makes irrelevant comments (OPP 7)
E-14. Comes to class unprepared (APD 7)
E-15. Does not participate in class (APD 8)
E-16. Does not respect others space (OPP 8)
I-1. Spends too much time alone (SW)
I-2. Complains about being sick or hurt (AD)
I-4. Avoids social interactions (SW)
I-5. Prefers to play alone (SW)
I-6. Does not respond to others' attempts to socialize (SW)
I-7. Worries about unimportant details (AD)
I-8. Complains of headaches or stomach aches (AD)
I-9. Appears unhappy or sad (AD)
I-10. Clings to adults (AD)
I-11. Acts nervous (AD)
I-12. Acts fearful (AD)
I-13. Does not stick up for self (SW)
I-14. Overly shy (SW)
I-15. Complains or whines (AD)
I-16. Does not participate in group activities (SW)
I-17. Makes self-depreciating comments (AD)
I-19. Cries or is weepy (AD)
I-23. Spends a lot of time worrying (AD)
I-24. Slow to warm up to new people (SW)

References

Achenbach, T. M., Andreas Becker, Manfred Döpfner, Einar Heiervang, Veit Roessner, Hans-Christoph Steinhausen, and Aribert Rothenberger. 2008. Multicultural assessment of child and adolescent psychopathology with ASEBA and SDQ instruments: Research findings, applications, and future directions. *Journal of Child Psychology and Psychiatry* 49: 251–75. [CrossRef] [PubMed]

Achenbach, T. M., Stephanie H. McConaughy, Masha Y. Ivanova, and Leslie A. Rescorla. 2011. *Manual for the ASEBA Brief Problem Monitor*. Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.

Anthony, Christopher J., James C. Di Perna, and Pui-Wa Lei. 2016. Maximizing Measurement Efficiency of Behavior Rating Scales Using Item Response Theory: An Example with the Social Skills Improvement System—Teacher Rating Scale. *Journal of School Psychology* 55: 57–69. [CrossRef] [PubMed]

Anthony, Christopher J., Kara M. Styck, Robert J. Volpe, and Christopher R. Robert. 2022. Using many-facet rasch measurement and generalizability theory to explore rater effects for direct behavior rating–multi-item scales. *School Psychology online first*. [CrossRef] [PubMed]

Aviles, Ann M., Tanya R. Anderson, and Erica R. Davila. 2006. Child and Adolescent Social-Emotional Development Within the Context of School. *Child and Adolescent Mental Health* 11: 32–39. [CrossRef] [PubMed]

Becker, Janine, Carolyn Schwartz, Renee N. Saris-Baglama, Mark Kosinski, and Jakob Bue Bjorner. 2007. Using Item Response Theory (IRT) For Developing and Evaluating the Pain Impact Questionnaire (PIQ-6™). *Pain Medicine* 8: 129–44. [CrossRef]

Berg, Juliette, Elizabeth Nolan, Nick Yoder, David Osher, and Amy Mart. 2019. Social-Emotional Competencies in Context: Using Social-Emotional Learning Frameworks to Build Educators' Understanding. *Measuring SEL* 2019: 1–13.

Breitenstein, Susan M., Carri Hill, and Deborah Gross. 2009. Understanding disruptive behavior problems in preschool children. *Journal of Pediatric Nursing* 24: 3–12. [CrossRef]

Briesch, Amy M., Hariharan Swaminathan, Megan Welsh, and Sandra M. Chafouleas. 2014. Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology* 52: 13–35. [CrossRef]

Briesch, Amy M., Tyler David Ferguson, Brian Daniels, Robert J. Volpe, and Adam B. Feinberg. 2017. Examining the Influence of Interval Length on the Dependability of Observational Estimates. *School Psychology Review* 46: 426–32. [CrossRef]

Bruhn, Allison Leigh, Suzanne Woods-Groves, and Sally Huddle. 2014. A Preliminary Investigation of Emotional and Behavioral Screening Practices in K–12 Schools. *Education and Treatment of Children* 37: 611–34. [CrossRef]

Burns, John R., and Ronald M. Rapee. 2019. School-Based Assessment of Mental Health Risk in Children: The Preliminary Development of the Child RADAR. *Child and Adolescent Mental Health* 24: 66–75. [CrossRef]

Casale, Gino, Robert J. Volpe, Brian Daniels, Thomas Hennemann, Amy M. Briesch, and Michael Grosche. 2018. Measurement Invariance of a Universal Behavioral Screener Across Samples from the USA and Germany. *European Journal of Psychological Assessment* 34: 87–100. [CrossRef]

Casale, Gino, Robert J. Volpe, Thomas Hennemann, Amy M. Briesch, Brian Daniels, and Michael Grosche. 2019. Konstruktvalidität Eines Universellen Screenings Zur Unterrichtsnahen Und Ökonomischen Diagnostik Herausfordernden Verhaltens Von Schüler_innen–Eine Multitrait-Multimethod-Analyse. [Construct Validity of a Universal Screener to Economically Assess Students' Behavior in the Classroom—A Multitrait-Multimethod-Analysis]. *Zeitschrift für Pädagogische Psychologie [German Journal of Educational Psychology]* 33: 17–31. [CrossRef]

Chiesi, Francesca, Kinga Morsanyi, Maria Anna Donati, and Caterina Primi. 2018. Applying Item Response Theory to Develop a Shortened Version of the Need for Cognition Scale. *Advances in Cognitive Psychology* 14: 75–86. [CrossRef]

Christensen, Karl Bang, Guido Makransky, and Mike Horton. 2017. Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Applied Psychological Measurement* 41: 178–94. [CrossRef] [PubMed]

Costello, E. Jane. 2016. Early detection and prevention of mental health problems: Developmental epidemiology and systems of support. *Journal of Clinical Child & Adolescent Psychology* 45: 710–17.

Daniels, Brian, Robert J. Volpe, Amy M. Briesch, and Gregory A. Fabiano. 2014. Development of a Problem-Focused Behavioral Screener Linked to Evidence-Based Intervention. *School Psychology Quarterly* 29: 438–51. [CrossRef] [PubMed]

Dineen, Jennifer N., Sandra M. Chafouleas, Amy M. Briesch, D. Betsy McCoach, Sarah D. Newton, and Dakota W. Cintron. 2022. Exploring Social, Emotional, and Behavioral Screening Approaches in U.S. Public School Districts. *American Educational Research Journal* 59: 146–79. [CrossRef]

Domitrovich, Celene E., Joseph A. Durlak, Katharine C. Staley, and Roger P. Weissberg. 2017. Social-Emotional Competence: An Essential Factor for Promoting Positive Adjustment and Reducing Risk in School Children. *Child Development* 88: 408–16. [CrossRef]

Durlak, Joseph A., Roger P. Weissberg, Allison B. Dymnicki, Rebecca D. Taylor, and Kriston B. Schellinger. 2011. The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. *Child Development* 82: 405–32. [CrossRef]

Dwyer, Sarah B., Jan M. Nicholson, and Diana Battistutta. 2006. Parent and teacher identification of children at risk of developing internalizing or externalizing mental health problems: A comparison of screening methods. *Prevention Science* 7: 343–57. [CrossRef] [PubMed]

Eklund, Katie, Tyler L. Renshaw, Erin Dowdy, Shane R. Jimerson, Shelley R. Hart, Camille N. Jones, and James Earhart. 2009. Early Identification of Behavioral and Emotional Problems in Youth: Universal Screening Versus Teacher-Referral Identification. *California School Psychologist* 14: 89–95. [CrossRef]

Forness, Steven R., Joanne Kim, and Hill M. Walker. 2012. Prevalence of Students with EBD: Impact on General Education. *Beyond Behavior* 21: 3–10.

Glover, Todd A., and Craig A. Albers. 2007. Considerations for Evaluating Universal Screening Assessments. *Journal of School Psychology* 45: 117–35. [CrossRef]

Goodman, R. 1997. The Strengths and Difficulties Questionnaire: A Research Note. *Journal of Child Psychology and Psychiatry* 38: 581–86. [CrossRef]

Halle, Tamara G., and Kristen E. Darling-Churchill. 2016. Review of Measures of Social and Emotional Development. *Journal of Applied Developmental Psychology* 45: 8–18. [CrossRef]

Hambleton, Ronald K. 2000. Emergence of Item Response Modeling in Instrument Development and Data Analysis. *Medical Care* 38: II60–II65. [CrossRef]

Hartman, Kelsey, Frank M. Gresham, and Shelby Byrd. 2017. Student internalizing and externalizing behavior screeners: Evidence for reliability, validity, and usability in elementary schools. *Behavioral Disorders* 42: 108–18. [CrossRef]

Kauffman, James M. 1999. How We Prevent the Prevention of Emotional and Behavioral Disorders. *Exceptional Children* 65: 448–68. [CrossRef]

Kendziora, Kimberly T. 2004. Early Intervention for Emotional and Behavioral Disorders. In *Handbook of Research in Emotional and Behavioral Disorders*. Edited by Robert B. Rutherford, Mary M. Quinn and Sarup R. Mathur. New York, NY: The Guilford Press, pp. 327–51.

Kilgus, Stephen P., Nathaniel P. von der Embse, Crystal N. Taylor, Michael P. Van Wie, and Wesley A. Sims. 2018. Diagnostic accuracy of a universal screening multiple gating procedure: A replication study. *School Psychology Quarterly* 33: 582. [CrossRef]

Korpershoek, Hanke, Truus Harms, Hester de Boer, Mechteld van Kuijk, and Simone Doolaard. 2016. A Meta-Analysis of the Effects of Classroom Management Strategies and Classroom Management Programs on Students' Academic, Behavioral, Emotional, and Motivational Outcomes. *Review of Educational Research* 86: 643–80. [CrossRef]

Kovess-Masfety, Viviane, Mathilde M. Husky, Katherine Keyes, Ava Hamilton, Ondine Pez, Adina Bitfoi, Mauro Giovanni Carta, Dietmar Goelitz, Rowella Kuijpers, Roy Otten, and et al. 2016. Comparing the Prevalence of Mental Health Problems in Children 6–11 Across Europe. *Social Psychiatry and Psychiatric Epidemiology* 51: 1093–103. [CrossRef]

Lane, Kathleen, Wendy Peia Oakes, Holly Mariah Menzies, and Kathryn A. Germer. 2014. Screening and identification approaches for detecting students at risk. In *Handbook of Evidence-Based Practices for Emotional and Behavioral Disorders: Applications in Schools*. Edited by Hill Walker and Frank M. Gresham. New York, NY: Guilford Press, pp. 129–51.

Langer, David A., Jeffrey J. Wood, Patricia A. Wood, Ann F. Garland, John Landsverk, and Richard L. Hough. 2015. Mental health service use in schools and non-school-based outpatient settings: Comparing predictors of service use. *School Mental Health* 7: 161–73.

Moore, Stephanie A., Erin Dowdy, Tameisha Hinton, Christine DiStefano, and Fred W. Greer. 2022. Moving Toward Implementation of Universal Mental Health Screening by Examining Attitudes Toward School-Based Practices. *Behavioral Disorders* 47: 166–75. [CrossRef] [PubMed]

Muraki, Eiji. 1997. A Generalized Partial Credit Model. In *Handbook of Modern Item Response Theory*. Edited by Wim J. Linden and Ronald K. Hambleton. New York, NY: Springer, pp. 153–64.

Petermann, Ulrike, and Franz Petermann. 2013. *Lehrereinschätzliste für Sozial- und Lernverhalten*, 2nd ed. Teacher Assessment Scale for Social and Learning Behavior. Göttingen: Hogrefe.

Polanczyk, Guilherme V., Giovanni A. Salum, Luisa S. Sugaya, Arthur Caye, and Luis A. Rohde. 2015. Annual Research Review: A Meta-Analysis of the Worldwide Prevalence of Mental Disorders in Children and Adolescents. *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 56: 345–65. [CrossRef] [PubMed]

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 11 October 2022).

Revelle, William. 2022. *psych: Procedures for Personality and Psychological Research*. Evanston: Northwestern University, Version 2.2.9. Available online: https://CRAN.R-project.org/package=psych (accessed on 11 September 2022).

Robitzsch, Alexander, Thomas Kiefer, and Margaret Wu. 2022. TAM: Test Analysis Modules. R package Version 4.1–4. Available online: https://CRAN.R-project.org/package=TAM (accessed on 26 September 2022).

Sklad, Marcin, René Diekstra, Monique de Ritter, Jehonathan Ben, and Carolien Gravesteijn. 2012. Effectiveness of School-Based Universal Social, Emotional, and Behavioral Programs: Do They Enhance Students' Development in the Area of Skill, Behavior, and Adjustment? *Psychology in the Schools* 49: 892–909. [CrossRef]

Song, Yoon Ah. 2019. A Comparative Study of IRT Models for Rater Effects and Double Scoring. Doctoral dissertation, The University of Iowa, Iowa, IA, USA.

Splett, Joni W., Marlene Garzona, Nicole Gibson, Daniela Wojtalewicz, Anthony Raborn, and Wendy M. Reinke. 2019. Teacher Recognition, Concern, and Referral of Children's Internalizing and Externalizing Behavior Problems. *School Mental Health: A Multidisciplinary Research and Practice Journal* 11: 228–39. [CrossRef]

Stiffler, Meghan C., and Bridget V. Dever. 2015. Multiple-gating and mental health screening. In *Mental Health Screening at School: Instrumentation, Implementation, and Critical Issues*. Contemporary Issues in Psychological Assessment. Edited by Meghan C. Stiffler and Bridget V. Dever. Cham: Springer International Publishing/Springer Nature, pp. 91–105. [CrossRef]

Volpe, Robert J., and Amy M. Briesch. 2018. Establishing evidence-based behavioral screening practices in US schools. *School Psychology Review* 47: 396–402. [CrossRef]

Volpe, Robert J., and Gregory A. Fabiano. 2013. *Daily Behavior Report Cards: An Evidence-Based System of Assessment and Intervention*. New York, NY: Guilford Press.

Volpe, Robert J., and Kenneth D. Gadow. 2010. Creating abbreviated rating scales to monitor classroom inattention-overactivity, aggression, and peer conflict: Reliability, validity, and treatment sensitivity. *School Psychology Review* 39: 350–63. [CrossRef]

Volpe, Robert J., Amy M. Briesch, and Sandra M. Chafouleas. 2010. Linking Screening for Emotional and Behavioral Problems to Problem-Solving Efforts: An Adaptive Model of Behavioral Assessment. *Assessment for Effective Intervention* 35: 240–44. [CrossRef]

Volpe, Robert J., Amy M. Briesch, and Kenneth D. Gadow. 2011. The efficiency of behavior rating scales to assess disruptive classroom behavior: Applying generalizability theory to streamline assessment. *Journal of School Psychology* 49: 131–55. [CrossRef] [PubMed]

Volpe, Robert J., Gino Casale, Changiz Mohiyeddini, Michael Grosche, Thomas Hennemann, Amy M. Briesch, and Brian Daniels. 2018. A Universal Behavioral Screener Linked to Personalized Classroom Interventions: Psychometric Characteristics in a Large Sample of German Schoolchildren. *Journal of School Psychology* 66: 25–40. [CrossRef] [PubMed]

Volpe, Robert J., Tat Shing Yeung, Gino Casale, Johanna Krull, Amy M. Briesch, and Thomas Hennemann. 2020. Evaluation of a German Language School-Based Universal Screening for Student Social, Emotional, and Behavioral Risk. *International Journal of School & Educational Psychology* 9: 10–20. [CrossRef]

Walker, Hill M., Jason W. Small, Herbert H. Severson, John R. Seeley, and Edward G. Feil. 2014. Multiple-Gating Approaches in Universal Screening Within School and Community Settings. In *Universal Screening in Educational Settings: Evidence-Based Decision Making for Schools*. Edited by Ryan J. Kettler, Todd A. Glover, Craig A. Albers and Kelly A. Feeney-Kettler. Washington, DC: American Psychological Association, pp. 47–75.

Whitcomb, Sara A., and Kenneth W. Merrell. 2013. *Behavioral, Social, and Emotional Assessment of Children and Adolescents*, 4th ed. New York, NY: Routledge.

Wilson, Mark. 2004. *Constructing Measures: An Item Response Modeling Approach*. Mahwah: Lawrence Erlbaum Associates.

Wood, Brandon J., and Faith Ellis. 2022. Universal Mental Health Screening Practices in Midwestern Schools: A Window of Opportunity for School Psychologist Leadership and Role Expansion? *Contemporary School Psychology* 2022: 1–11. [CrossRef] [PubMed]

Wu, Margaret. 2017. Some IRT-Based Analyses for Interpreting Rater Effects. *Psychological Test and Assessment Modeling* 59: 453–70.

Journal of
Intelligence

MDPI

Article

Examining Humans' Problem-Solving Styles in Technology-Rich Environments Using Log File Data

Yizhu Gao [1,*], Xiaoming Zhai [2], Okan Bulut [1], Ying Cui [1] and Xiaojian Sun [3]

[1] Department of Educational Psychology, University of Alberta, Edmonton, AB T6G 2G5, Canada; bulut@ualberta.ca (O.B.); yc@ualberta.ca (Y.C.)
[2] Department of Mathematics, Science, and Social Studies Education, University of Georgia, Athens, GA 30602, USA; Xiaoming.Zhai@uga.edu
[3] School of Mathematics and Statistics, Southwest University, Chongqing 400715, China; sunxiaojian@swu.edu.cn
* Correspondence: yizhu@ualberta.ca

Abstract: This study investigated how one's problem-solving style impacts his/her problem-solving performance in technology-rich environments. Drawing upon experiential learning theory, we extracted two behavioral indicators (i.e., planning duration for problem solving and human–computer interaction frequency) to model problem-solving styles in technology-rich environments. We employed an existing data set in which 7516 participants responded to 14 technology-based tasks of the Programme for the International Assessment of Adult Competencies (PIAAC) 2012. Clustering analyses revealed three problem-solving styles: *Acting* indicates a preference for active explorations; *Reflecting* represents a tendency to observe; and *Shirking* shows an inclination toward scarce tryouts and few observations. Explanatory item response modeling analyses disclosed that individuals with the *Acting* style outperformed those with the *Reflecting* or the *Shirking* style, and this superiority persisted across tasks with different difficulties.

Keywords: problem-solving style technology-rich environments; experiential learning theory; *k*-means clustering; explanatory item response modeling; log file data

Citation: Gao, Yizhu, Xiaoming Zhai, Okan Bulut, Ying Cui, and Xiaojian Sun. 2022. Examining Humans' Problem-Solving Styles in Technology-Rich Environments Using Log File Data. *Journal of Intelligence* 10: 38. https://doi.org/10.3390/jintelligence10030038

Received: 20 April 2022
Accepted: 25 June 2022
Published: 30 June 2022

1. Introduction

As information and communication technologies rapidly integrate into people's everyday lives, the importance of being able to use technological tools to solve problems continues to grow in recent years (Hämäläinen et al. 2015; Koehler et al. 2017; Zheng et al. 2017). As highlighted by Iñiguez-Berrozpe and Boeren (2020), being insufficient to solve technology-based problems can exclude one from the labor market. This has been particularly true when people felt challenged to use computers or other digital devices to perform work-related activities (Hämäläinen et al. 2015; Ibieta et al. 2019; Nygren et al. 2019; Tatnall 2014). Nonetheless, a huge amount of people seem to have insufficient problem-solving performance in technology-rich environments (TRE). As pointed out by Nygren et al. (2019), more than 50% of European aged 16–64 years old were deficient in coping with practical tasks in TRE (e.g., communicating with others by email). Notably, TRE incorporate diverse, versatile, and constantly evolving digital technologies, leading to difficulties in being operated expertly. Considering feasibility reasons, TRE in the present study are limited to settings involving the most common digital technologies (Nygren et al. 2019): computers (e.g., spreadsheet) and Internet-based services (e.g., web browser). To boost the use of digital technologies, a bulk of research has investigated factors that might affect humans' problem-solving performance in TRE (e.g., Liao et al. 2019; Millar et al. 2020; Nygren et al. 2019; Ulitzsch et al. 2021). Among those findings, problem-solving style was regarded as one of the most prominent factors (e.g., Koć-Januchta et al. 2020; Lewis and Smith 2008; Treffinger et al. 2008).

Problem-solving style describes pervasive aspects of individuals' natural dispositions toward problem solving. According to Selby et al. (2004, p. 222), problem-solving styles are "consistent individual differences in the ways people prefer to plan and carry out generating and focusing activities, in order to gain clarity, produce ideas, and prepare for action". This broadly accepted definition indicates that problem-solving style derives from one's distinguishable behavioral pattern (e.g., He et al. 2021; Ulitzsch et al. 2021). In this regard, problem-solving styles in TRE reflect individuals' dispositions regarding how they are inclined to interact with surrounding technology environments. Implicit tendencies, in turn, can be partially explicated by behavioral indicators recorded in computer-generated log files, such as timestamps, clicks, and sequence of actions (Bunderson et al. 1989; Eichmann et al. 2019; Oshima and Hoppe 2021). In other words, a critical empirical avenue to profiling an individual's problem-solving style in TRE is to analyze log file data collected in computer-based problem-solving assessments.

This study analyzed log file data of the Programme for the International Assessment of Adult Competencies (PIAAC) 2012 to unpack problem-solving styles in TRE and examined how problem-solving styles were associated with participants' performance on TRE-related tasks. In PIAAC 2012, a total of 14 tasks were administered to assess participants' problem-solving competencies in TRE, all of which simulate real-world problems that adults likely encounter when using computers and Internet-based technologies. The data from assessment tasks provide rich information, such as performance and behavioral information. However, abstracting the useful information from the log files is challenging because multiple variables with manifold types are embedded in the data structure (Han et al. 2019). To overcome this challenge, we first applied clustering techniques to multiple behavioral indicators derived from the 14 tasks, thereby partitioning participants into discrepant clusters. Each cluster was further analyzed and its specific problem-solving style was identified according to behavioral indicators. Finally, we examined how the personal features (i.e., problem-solving style) and their interaction with task features (i.e., task difficulty level) account for participants' task performance by explanatory item response modeling (EIRM; De Boeck and Wilson 2004).

1.1. Problem-Solving Styles in TRE

In this study, the problem-solving style in TRE is conceptualized and operationalized as the consistent individual behavior in planning and carrying out problem-solving activities in surrounding technology environments (Isaksen et al. 2016; Selby et al. 2004; Treffinger et al. 2008). Despite the importance and the pervasiveness of problem-solving styles, few pertinent theories have been put forward in this area. A potential theory that may enlighten our understanding of problem-solving styles in TRE is experiential learning theory (Kolb 2015). Experiential learning theory emphasizes the central role of experience in human learning and development processes and has been widely accepted as a useful framework for educational innovations (Botelho et al. 2016; Koivisto et al. 2017; Morris 2020). In his seminal works, Kolb (2015) suggests four types of learning modes to portray individuals' learning preferences as a combination of grasping and transforming experiences: if individuals prefer an abstract grasping of information from experiences, their inclined learning mode is abstract conceptualization (AC); in contrast, if individuals prefer highly contextualized and hands-on experiences, their learning mode is known as concrete experience (CE); if individuals prefer to act upon the grasped information, their preference of transforming experience is active experimentation (AE); otherwise, their preferred way may be reflective observation (RO). Thereafter, much research has studied learning styles based on individuals' relative preferences for the four learning modes and agrees upon a nine-style typology (e.g., Eickmann et al. 2004; Kolb and Kolb 2005a; Sharma and Kolb 2010). Specifically, four learning styles emphasize one of the four learning modes; another four represent learning style types that emphasize two learning modes; one learning style type balances all the four learning modes. For example, learning styles of *Acting* and *Reflecting* correspond to learning modes of AE and RO, respectively. Individuals with the

Acting style usually possess highly developed action skills while utilizing little reflection (AE). In contrast, those with the *Reflecting* style spend much time buried in their thoughts, but have trouble putting plans into action (RO).

Learning modes are highly associated with problem-solving styles. There is an emerging consensus that learning interacts with and contributes to ongoing problem-solving processes (Ifenthaler 2012; Wang and Chiew 2010). Research has indicated that problem solving is not only a knowledge application process but also a knowledge acquisition and accumulation process. In this respect, humans' learning modes along with exploring problem environments can be part of problem-solving styles (Kim and Hannafin 2011). For example, Romero et al. (1992) developed the Problem-Solving Style Questionnaire based on a hypothesized problem-solving process in which the four learning modes (i.e., CE, RO, AC, and AE) are involved. Besides the close conceptual connections between learning modes and problem-solving styles, learning modes are increasingly incorporated into designing technology-enhanced learning environments given their capability to describe users' online learning styles. For example, Richmond and Cummings (2005) discussed the integration of learning modes with online distance education and suggested that learning modes should be considered for instructional design to ensure high-quality online courses and to achieve positive student outcomes. In addition, an earlier study by Bontchev et al. (2018) has demonstrated the usefulness of learning modes in enlightening humans' styles in game-based problem solving. Therefore, learning modes can potentially inform the types of problem-solving styles in TRE.

1.2. Acting and Reflecting Styles

Among learning styles portrayed in a two-dimensional learning space defined by AC-CE and AE-RO, the *Acting* and *Reflecting* styles are particularly representative of individual interactive modes in TRE. For example, Hung et al. (2016) took the *Acting* and *Reflecting* styles into account when they provided adaptive suggestions to optimize problem-solving performance in computer-based environments. Bontchev et al. (2018) investigated problem-solving styles within educational computer games, which correspond to the *Acting* and *Reflecting* styles. These studies confirmed that the *Acting* and *Reflecting* styles are feasible to describe problem-solving styles in TRE.

A distinctive feature of the *Acting* style is the strong motivation for goal-directed actions that integrate people and objects (Kolb and Kolb 2005b). Individuals with the *Acting* style prefer to work and try objects out (Hung et al. 2016). Within TRE, individuals with the *Acting* style habitually perform actions quickly and frequently, which implies their intuitive readiness to act. In contrast, the *Reflecting* style is characterized by the tendency to connect experience and ideas through sustained reflections (Kolb and Kolb 2005b). Individuals with the *Reflecting* style prefer to evaluate and think about objects (Hung et al. 2016). When interacting with objects in TRE, they need time to observe and establish the meaning of available operations in technological environments. They watch patiently rather than automatic reaction and wait to act until certain of their intention.

In addition to their suitability for describing problem-solving styles in TRE, evidence shows that the *Acting* and *Reflecting* styles are relevant to problem-solving performance. For example, Kolb and Fry (1975, p. 54) suggested that a behaviorally complex learning environment distinguished by "environmental responses contingent upon self-initiated action" emphasizes actively applying knowledge or skills to practical problems, and thus better supports the learning mode of AE. Following this view, individuals with the *Acting* style are supposed to have better performance in TRE-related tasks than those with the *Reflecting* style who have deficiencies in AE. However, this theoretical assumption needs to be empirically examined.

Furthermore, it is crucial to consider the role of problem characteristics (e.g., problem type or problem difficulty) in the relationship between individuals' problem-solving styles and their performance in problem solving. As stated by Treffinger et al. (2008), an individual's preference for a certain problem-solving style can influence his or her behavior in

finding, defining, and solving problems. That is, a certain problem-solving style can either hamper or facilitate problem-solving performance, depending on some characteristics of problems. For example, Treffinger et al. (2008) found that individuals with the explorer style deal well with ill-defined and ambiguous problems, while individuals with the developer style are adept at handling well-defined problems. Thus, studies need to examine the role of problem characteristics when investigating the impact of problem-solving styles on problem-solving performance.

1.3. Behavioral Indicators of Acting and Reflecting Styles in TRE

To examine the feasibility of the *Acting* and *Reflecting* styles in describing problem-solving behaviors in TRE, two behavioral indicators were abstracted from log files: duration of planning period at the beginning of the problem-solving process and interaction frequency during the entire problem-solving process. For simplicity, the two behavioral indicators were abbreviated as planning duration and interaction frequency, respectively. Planning duration denotes the period from the time that a task starts to the point that people take their first action to perform the task. It is also called first move latency (e.g., Albert and Steinberg 2011; Eichmann et al. 2019) or timing of the first action (e.g., Goldhammer et al. 2016; Liao et al. 2019). In this study, the term "planning duration" is used to emphasize people's thinking and reflection on the problem at hand (Albert and Steinberg 2011). Interaction frequency indicates how frequently people interact with a task during the period from the first action to the end of the task.

The two indicators formulate a two-dimensional space that could portray individuals' problem-solving behaviors. Specifically, based on previous research (e.g., Eickmann et al. 2004; Hung et al. 2016; Kolb and Kolb 2005a), individuals with the *Acting* style prefer to act on tasks with multiple trials while seldom reflecting on their behaviors during the course. They perform like experimentalists. In contrast, those with the *Reflecting* style prefer to fully reflect on situations instead of taking concrete actions. They tend to be theoreticians. During problem solving in TRE, individuals with the *Acting* style usually spend less time on planning, but interact more with objects in comparison with those with the *Reflecting* style who spare more time for planning, but execute tasks less.

Although the role of planning duration and interaction frequency in problem solving has been widely studied previously (Albert and Steinberg 2011; Eichmann et al. 2019; Greiff et al. 2016), no study has explored how these two measures together inform individual problem-solving styles in TRE. Albert and Steinberg (2011) found that planning time, which reflects self-regulatory control, strongly and positively predicted outcomes of problem solving. However, a longer time of first-move latency may not necessarily indicate participants as being more thoughtful. Instead, participants may merely feel confused about problems (Zoanetti and Griffin 2014). In fact, interaction frequency could cooperate with planning duration in inferring participants' inclination toward problem solving in TRE (Eichmann et al. 2019). For example, a thoughtful individual would not only spend more time planning at the beginning but also have relatively fewer tryouts during the problem-solving process, indicating their accurate reasoning and confident judgments.

1.4. Current Study

Given the limited volume of research on humans' problem-solving styles in TRE, this study first examined *Acting* and *Reflecting* styles in TRE using two indicators: planning duration and interaction frequency. We then compared different problem-solving styles to identify the most desirable one for solving technology-based problems. Finally, we examined how task difficulty moderates the relationship between individual task performance and individual problem-solving styles. The study answers three research questions:

1. Did participants demonstrate *Acting* or *Reflecting* problem-solving styles when solving problems in TRE?
2. If so, which problem-solving style better favors participants' performance?

3. How did task difficulty moderate the relationship between participants' problem-solving styles and their performance on TRE-related tasks?

2. Materials and Methods

2.1. Participants

We employed existing data from the PIAAC 2012 conducted by the Organisation for Economic Co-operation and Development (OECD). In total 81,744 participants aged 16 to 65 from 17 countries participated in the PIAAC test (Organisation for Economic Co-operation and Development (OECD) 2013). The participants were randomly assigned to two of the three cognitive modules, each of which comprised either literacy, numeracy, or problem-solving in TRE (PSTRE) tasks (Organisation for Economic Co-operation and Development (OECD) 2013). We analyzed 10,806 participants who responded to two PSTRE modules from 14 of the 17 countries, as data from three countries (i.e., France, Italy, and Spain) were not available. We cleaned the invalid data as some participants merely pressed the next button without responding to the questions. Participants with outliers in terms of three variables (i.e., the timing of the first action, the total number of interactions, and the duration of the entire problem-solving process) were also excluded. Outliers were identified by examining whether values lay outside of three standard deviations of the average value. Eventually, N = 7516 participants with an average age of 36.29 years (SD = 13.62) were included in the analysis, of which 47.90% were male. The demographic information of participants included in the study was presented in Table 1 by country.

Table 1. Demographic Information of Participants in the Present Study.

Country	*N*	Gender		Age	
		Male	Female	Average	*SD*
Austria	414	227	187	NA [1]	NA [1]
Belgium	503	255	248	37.29	13.74
Denmark	684	316	368	42.31	14.44
Estonia	628	283	345	35.73	13.21
Finland	501	264	237	37.44	13.22
Germany	420	206	214	NA [1]	NA [1]
Ireland	492	236	256	36.94	11.77
Republic of Korea	465	226	239	33.71	11.84
Netherlands	521	242	279	39.11	14.49
Norway	496	253	243	37.96	13.54
Poland	711	352	359	26.25	9.90
Slovakia	383	197	186	33.57	12.99
United Kingdom	869	338	531	38.51	12.91
United States	429	205	224	NA [1]	NA [1]

[1] NA indicates there is no available information.

2.2. Instruments

The PSTRE domain aims to measure "abilities to solve problems for personal, work and civic purposes by setting up appropriate goals and plans and accessing and making use of information through computers and computer networks" (Organisation for Economic Co-operation and Development (OECD) 2013, p. 56). Accordingly, 14 computerized tasks were developed to mimic real-life problems that adults are likely to encounter while using computers and Internet-based technologies (Organisation for Economic Co-operation and Development (OECD) 2019). Organisation for Economic Co-operation and Development (OECD) (2012, p. 48) defined three core dimensions when developing the 14 tasks. The first dimension is problem circumstances that trigger a person's curiosity about problem solving and determine actions required to be taken to solve problems. The second is technologies through which problem solving is conducted, such as computer devices, applications, and functionalities. The third dimension is cognitive processes underlying problem solving (e.g., goal setting and reasoning). These three dimensions played an intertwined role in

distinguishing participants' proficiency levels in PSTRE. For example, the "Job Search" task (see Figure 1) creates a scenario in which participants assume that they are taking the role of job seekers. Participants click on links or forward/back icons and then bookmark as many web pages as possible. If participants solve this task, it is assumed that they can identify problem goals and operate technology applications. Three proficiency levels of PSTRE in total were distinguished in the PIAAC 2012 and 14 tasks were distributed over three difficulty levels (Organisation for Economic Co-operation and Development (OECD) 2019). More challenging tasks have higher difficulty levels: three, seven, and four tasks were at difficulty levels 1, 2, and 3 correspondingly. All participants finished each PSTRE module within 30 min. The order of tasks within each module and that of the modules were always the same. Participants were not allowed to return to a former task after finishing it.

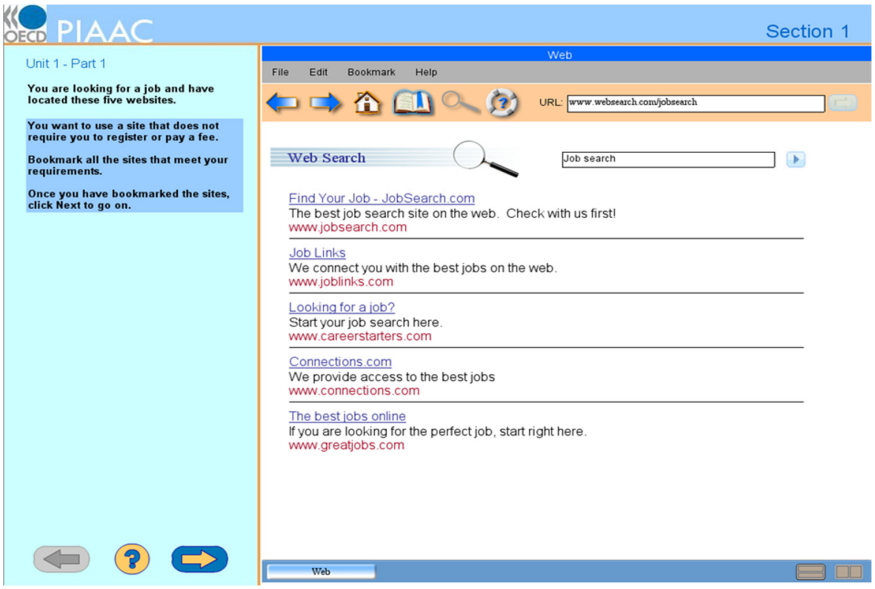

Figure 1. This is an exemplary problem-solving item in TRE. From Job Search Part I, by (Organisation for Economic Co-operation and Development (OECD) n.d.) (https://piaac-logdata.tba-hosting.de/public/problemsolving/JobSearchPart1/pages/jsp1-home.html) (accessed on 11 August 2021).

2.3. Scoring

2.3.1. Task Rubric and Scoring

According to the PIAAC technical report (Organisation for Economic Co-operation and Development (OECD) 2016), it is based on predefined scoring rubrics to grade participants' responses. As shown in Table 2, task scores are of mixed formats: eight tasks were dichotomously scored (i.e., correct, incorrect), and six tasks were polytomously scored (i.e., full, partial, no credit).

Table 2. Scoring Types and Scores of the 14 Tasks.

Task	Type	Scores
1	P	0, 1, 2, 3
2	D	0, 1
3	P	0, 1, 2, 3
4	D	0, 1
5	P	0, 1, 2, 3
6	D	0, 1
7	D	0, 1
8	D	0, 1
9	P	0, 1, 2, 3
10	D	0, 1
11	D	0, 1
12	P	0, 1, 2
13	D	0, 1
14	P	0, 1, 2, 3

Note: D indicates the task is dichotomously scored. P denotes the task is polytomously scored.

2.3.2. Behavioral Indicators Scoring

To address our research questions, planning duration and interaction frequency were extracted as behavioral indicators from log file data for the 14 PSTRE tasks in the PIAAC 2012. We used the time between participants' view of the task and their first interaction as a measure of planning duration for one task. Thus, we had 14 measures of planning duration for each participant. Table 3 shows the descriptive statistics of these measures ranging from 0 to 16.28 min. The mean planning duration ranges from 0.26 min (*SD* = 0.19) to 0.82 min (*SD* = 0.49) for the 14 tasks. Planning durations of all tasks are almost normally distributed based on skewness values ranging from 0.72 to 1.90 (George and Mallery 2010) except for the eighth task with a skewness value of 11.72. The extremely long planning duration (16.28 min) may explain its highly skewed distribution.

Table 3. Descriptive Statistics of Planning Duration Indicator for 14 Tasks.

Task	Planning Duration (minutes)				
	M	***SD***	**Min**	**Max**	**Skewness**
1	0.56	0.34	0.00	2.51	1.52
2	0.48	0.28	0.00	1.68	0.72
3	0.38	0.25	0.00	1.75	1.10
4	0.72	0.49	0.00	5.47	1.70
5	0.57	0.56	0.00	3.86	1.90
6	0.82	0.49	0.00	2.96	0.95
7	0.33	0.25	0.00	1.39	1.03
8	0.52	0.38	0.00	16.28	11.72
9	0.26	0.19	0.00	1.16	1.13
10	0.43	0.28	0.00	1.65	0.90
11	0.79	0.62	0.00	3.58	1.58
12	0.54	0.37	0.00	2.03	0.78
13	0.55	0.29	0.00	1.90	0.89
14	0.39	0.24	0.00	1.42	0.80

For the behavioral indicator of interaction frequency, we calculated the ratio of the total number of human–computer interactions to the overall timing of interactions. The ratio was used because it normalizes the number of interactions for the timing. In addition, the ratio corresponds to core features that can distinguish different problem-solving styles effectively. The Appendix A displays a sample log data file that records sequences of actions undertaken by one participant of the PIAAC 2012. The log data file contains four variables associated with the problem-solving process in TRE. The "Item Name" variable indicates which task it is. Both the "Event Name" and "Event Type" variables explain behavioral events, which may be either system-generated (e.g., START, NEXT_ITEM, and END) or respondent-generated (e.g., CONFIRMATION_OPENED, MAIL_VIEWED, FOLDER_VIEWED). The "Timestamp" variable is the behavioral event time for the task given in milliseconds since the beginning of the assessment. We can infer that the respondent spent 0.24 min planning solutions and 2.94 min interacting with the task. Note that the overall timing of interactions is the duration from the first event to the end of the task (i.e., 2.94 min) instead of the overall timing of solving the problem (i.e., 3.18 min). Given that the total number of interactions was 45, the interaction frequency for this participant on the first task was 15.31 times/min. Similarly, we had 14 measures of interaction frequency for each respondent. As presented in Table 4, the mean interaction frequency ranged from 5.56 times/minute (*SD* = 3.30) to 18.53 times/minute (*SD* = 9.43). The skewness values show that the interaction frequencies for all tasks are normally distributed (George and Mallery 2010). It should be noted that the values of planning duration and interaction frequency did not share a common measurement scale. We thus rescale both variables using their ranges to compensate for the effect that different variations of planning duration and interaction frequency had on the following analysis (i.e., *k*-means clustering, (Henry et al. 2005)) results.

Table 4. Descriptive Statistics of Interaction Frequency Indicator for the 14 Tasks.

Task	Interaction Frequency (times/minute)				
	M	***SD***	**Min**	**Max**	**Skewness**
1	18.53	9.43	0.00	103.65	0.19
2	16.46	8.03	0.00	42.09	−0.30
3	11.25	6.42	0.00	34.55	0.25
4	8.27	5.74	0.00	28.85	0.99
5	10.87	9.45	0.00	86.26	1.29
6	5.56	3.30	0.00	20.19	1.30
7	6.36	3.97	0.00	20.67	0.60
8	11.48	4.96	0.00	27.38	−0.28
9	17.11	10.59	0.00	58.27	0.05
10	10.96	6.67	0.00	33.27	0.47
11	18.25	10.15	0.00	50.40	0.31
12	6.75	5.12	0.00	25.43	0.72
13	8.21	3.56	0.00	19.18	−0.03
14	12.85	7.08	0.00	46.10	0.45

2.4. Data Analysis

We first conducted *k*-means clustering with planning durations and interaction frequencies to categorize participants into different problem-solving styles groups. *k*-means clustering is one of the simplest learning algorithms for sample clustering. Using *k*-means clustering, one must first fix prior k-centroids and then assign each observation to the cluster associated with its nearest centroid (Jyoti and Singh 2011). We chose this algorithm for two reasons: first, the results of *k*-means clustering analysis are feasible to interpret because clusters can be distinguished by examining what respondents in each cluster have in common regarding their behavioral patterns; second, *k*-means clustering is efficient in terms of running-time even with a large number of participants and variables, which renders applications in large-scale assessments likely (He et al. 2019). One challenge to *k*-means clustering is to figure out the number of clusters in advance. We applied the average silhouette method to determine the optimal number of clusters (e.g., Kaufman and Rousseeuw 1990). Specifically, the average silhouette method calibrated the silhouette width to measure the difference between within-cluster distances and between-cluster distances. Kodinariya and Makwana (2013) compared six methods to automatically generate the optimal number of clusters, among which the average silhouette method had been recommended because it best improved the validation of the analysis results (Kaufman and Rousseeuw 1990). We thus employed the largest average silhouette width over different *k*s to identify the best number of clusters. Additionally, we used the NbClust method (Charrad et al. 2014) to validate the result from the average silhouette method. The NbClust method aims to gather all available indices of a data set (i.e., 30 indices), as presented by Charrad et al. (2014), to generate the optimal number of clusters. Using different combinations of cluster numbers, distance measures, and clustering methods, the NbClust method outputs a consensus on the best number of clusters for the data set.

k-means clustering employing the average silhouette method was first implemented using the package *factoextra* (Kassambara and Mundt 2020) in R (R Core Team 2022). We then used the *NbClust* package to validate the number of clusters from the average silhouette method. Next, the average scores on planning duration (i.e., 14 indicators) and interaction frequency (i.e., 14 indicators) were compared across clusters by one-way analysis of variance (ANOVA) separately to verify *Acting*/*Reflecting* styles in TRE, which was conducted using the *dplyr* package (Hadley Wickham et al. 2021) in R (R Core Team 2022).

EIRM was finally applied to understand the association between participants' problem-solving styles derived from the *k*-means clustering analysis and their performance on PSTRE and how consistent the association was across multiple item difficulty levels. Unlike traditional item response theory models that solely focus on the difficulty levels of individual items, EIRM allows task-level and person-level features as well as their interactions to be

incorporated into measurement models in order to explain the variation in task difficulties (De Boeck and Wilson 2004). This study employed a series of EIRM analyses, in which individuals' problem-solving styles identified by the *k*-means clustering were the person-level predictors, and task difficulty levels were the task-level predictors of participants' likelihood of completing the tasks correctly. We compared model fit indices and model variable coefficients to identify the most desired problem-solving style in TRE for participants. All EIRM analyses were implemented using the package *eirm* (Bulut 2021; Bulut et al. 2021) within the R computing environment (R Core Team 2022). Tasks with varying numbers of response categories were handled by the *polyreformat* function of the *eirm* package. Specifically, the *polyreformat* function transforms dichotomous and polytomous responses into a series of dummy-coded responses (Bulut et al. 2021). Figure 2 demonstrates how polytomous (i.e., task 1) and dichotomous response categories (i.e., task 2) are dichotomized in the new data set. For example, if a respondent had the response category of 3 for task 1, then the dummy-coded responses for this polytomous response would be 1 for 2–3 and missing (i.e., NA) for 0–1 and 1–2. If the respondent had the response category of 1 for task 2, then the dummy-coded responses for this dichotomous response would be 1 for 0–1, 0 for 1–2, and missing (i.e., NA) for 2–3. This series of dummy-coded responses can be performed with EIRM analyses together.

Person ID	Task ID	Response
1	1	3
1	2	1

⇨

Person ID	Task ID	Category	Response
1	1	0–1	NA
		1–2	NA
		2–3	1
		0–1	1
1	2	1–2	0
		2–3	NA

Figure 2. Examples of how polytomous and dichotomous responses are defined as pseudo-dichotomous responses.

3. Results

3.1. Are Acting and Reflecting Styles Applicable to Describe Problem-Solving Styles in TRE by Examining Planning Duration and Interaction Frequency?

We first used the average silhouette method to find the optimal number of clusters for the rescaled data. Figure 3 depicts the relationship between the average silhouette width and the cluster number ranging from one to ten. The three-cluster solution had the greatest silhouette width, suggesting that participants should be clustered into three groups based on their planning duration and interaction frequency on the 14 PSTRE tasks.

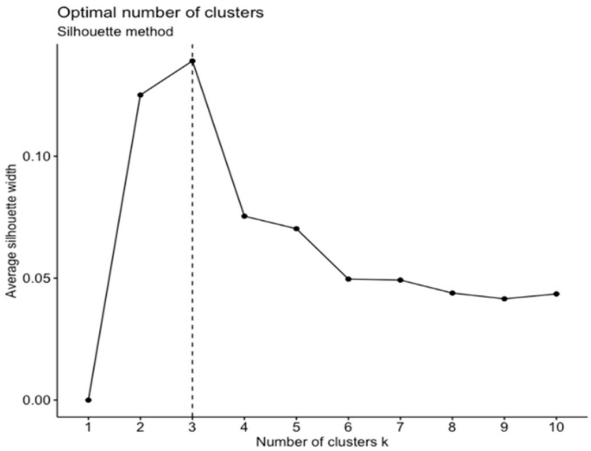

Figure 3. The optimal number of clusters by the average silhouette method for the two behavioral indicators.

To validate the three-cluster solution, we employed the NbClust method to generate a consensus on the optimal number of clusters for the data set. Figure 4 showed that the three-cluster solution was the one that was supported by most indices (i.e., 17).

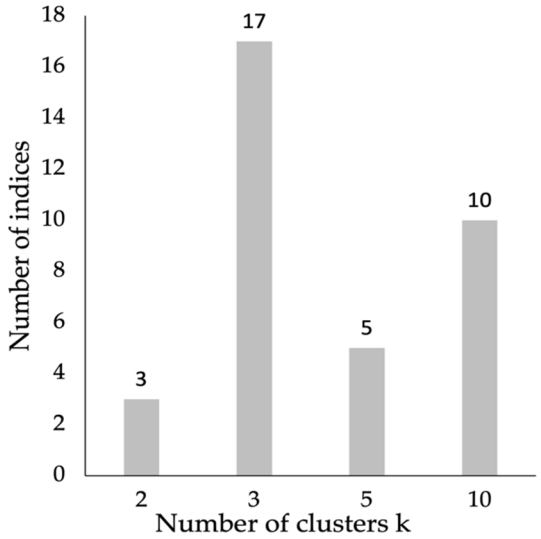

Figure 4. The optimal number of clusters suggested by the majority rule of the *NbClust* package for the two behavioral indicators.

To understand behavioral profiles for the three clusters, rescaled scores on planning duration and interaction frequency across the three clusters were shown in Figure 5. The larger the values were, the longer the planning duration or the higher interaction frequency that participants initiated. The mean rescaled scores on planning duration are 0.45 (SD = 0.06), −0.22 (SD = 0.08), and −0.59 (SD = 0.33) and the mean rescaled scores on interaction frequency are −0.24 (SD = 0.08), 0.46 (SD = 0.10), and −0.92 (SD = 0.27). Cluster 1 suggests the highest rescaled score on planning duration, but a lower rescaled score on interaction frequency, indicating that members of this cluster spent a particularly long time in action planning and did not devote much to the interaction with technology-based problems. In contrast, cluster 2 indicates the highest rescaled score on interaction frequency, but a lower rescaled score on planning duration, revealing that participants spent less time on setting up plans while actively interacting with TRE. Unlike clusters 1 and 2, cluster 3 suggests the lowest rescaled scores of both planning duration and interaction frequency. That is, respondents in cluster 3 barely spent time making plans before the operations that followed, and they were less frequently interacting with problem-solving tasks to solve problems.

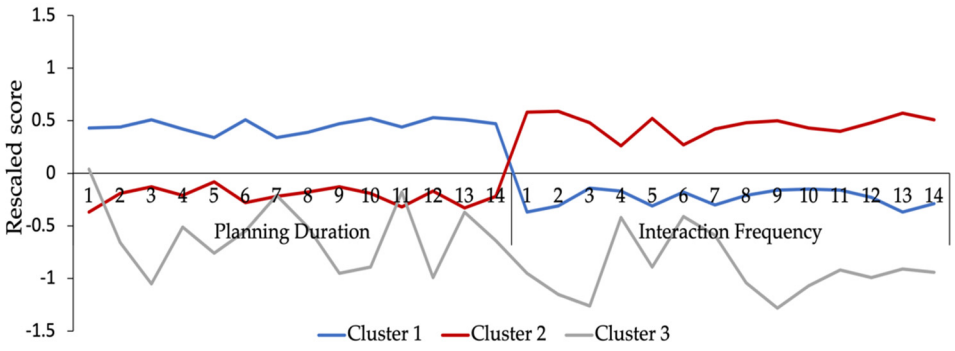

Figure 5. Behavioral profiles of the three clusters on the two behavioral indicators.

As shown in Table 5, of the participants, 2993 (39.82%), 3522 (46.86%), and 1001 (13.32%) were in clusters 1, 2, and 3, respectively. The mean values of planning dura-

tion and interaction frequency of the three clusters were also presented in Table 5. That is, solvers' planning duration for each PSTRE task was found to be 41.06 s for cluster 1 and decreased progressively to 26.70 and 19.50 s for clusters 2 and 3. The magnitude of interaction frequency for cluster 3 (5.14 times/min) was found to be lowest in comparison with cluster 1 (10.04 times/min) and cluster 2 (14.84 times/min). Two one-way ANOVAs were performed with solvers' clusters as the independent variable. Results indicated that differences in both behavioral indicators were significant across the three clusters, $F(2, 7513) = 4401$, $p < 0.001$, eta-squared = 0.540 and $F(2, 7513) = 7609$, $p < 0.001$, eta-squared = 0.670. Post hoc comparisons using the Tukey HSD method indicated that the planning duration of cluster 1 was the longest and the interaction frequency of cluster 2 was the highest among the three clusters. Thus, the behavioral patterns of clusters 1 and 2 were consistent with how individuals with *Reflecting* and *Acting* styles are expected to perform in TRE. We defined the problem-solving style of Cluster 3 as *Shirking* given its shortest planning duration and lowest interaction frequency.

Table 5. Summary of Two Behavioral Indicators of Each PSTRE Task for Three Clusters.

Cluster ID	N	Planning Duration (s)	Interaction Frequency (times/min)
1	2993	41.06	10.04
2	3522	26.70	14.84
3	1001	19.50	5.14

3.2. How Problem-Solving Styles Are Associated with Participants' Performance in PSTRE and How Does Task Difficulty Level Moderate Their Relationship?

To understand how task difficulty levels moderate the relationship between identified problem-solving styles in TRE and individual problem-solving performance, we conducted a series of EIRM analyses.

Model 0 represents the baseline model in which the only predictor was task difficulty levels at the task level. Difficulty scores of the 14 tasks reported by Organisation for Economic Co-operation and Development (OECD) (2019) were presented in Appendix B. We noted that tasks at the same difficulty level have close difficulty scores, while tasks at different difficulty levels differ greatly in their difficulty scores. The average difficulty score of tasks at difficulty level 2 (i.e., 311.7) lay outside of three standard deviations of the average difficulty score of tasks at difficulty level 1 (i.e., 274.0). It is the same when comparing tasks at difficulty level 3 with those at difficulty level 2. These pieces of information can corroborate Model 0. Model 1, as compared to Model 0, includes problem-solving styles as an additional predictor at the personal level. Lastly, Model 2 further incorporated the interaction between task difficulty and problem-solving style. The estimated parameters of Models 0, 1, and 2 are shown in Table 6. The baseline model (Model 0) shows that the estimated coefficients for task difficulty levels (TDL) are aligned with the PIAAC's categorization of task difficulty, where level 1 represents the easiest tasks ($b = -0.53$) and level 3 indicates the hardest tasks ($b = 1.92$). The next model, Model 1, compared the three clusters with different problem-solving styles: when compared with the *Reflecting* group (reference category), participants with the problem-solving style of *Shirking* were less likely to solve PSTRE tasks correctly ($OR = 0.17$; 83% less likely), whereas participants with the problem-solving style of *Acting* had a much higher chance of conducting the PSTRE tasks correctly ($OR = 1.58$; 58% more likely). The final model, Model 2, included two-way interactions between problem-solving styles and task difficulty levels. The interaction effects were statistically significant, but very small in magnitude, suggesting that task difficulty did not strongly moderate the relationship between problem-solving styles and participants' likelihood of solving TRE-related tasks. To directly compare the *Shirking* and the *Acting* group, we built another model (i.e., Model 1_Acting) including problem-solving styles as a predictor at the personal level and task difficulty levels as a predictor at the task level. Model 1_Acting is different from the current Model 1 because the control group in Model 1_Acting is *Acting* rather than *Reflecting*. We thus obtained the contrast between the

Shirking and the *Acting* style: participants with the problem-solving style of *Acting* were more likely to solve PSTRE tasks correctly in comparison with those with the *Shirking* style ($z = 63.70$, $p < 0.001$). Given that Model 1_Acting was built to compare the *Shirking* and the *Acting* style, we did not include the results of Model 1_Acting in Table 6 to keep EIRM analysis results in their current flow.

Table 6. A summary of EIRM results for Model 0, Model 1, and Model 2.

	Model 0				**Model 1**				**Model 2**			
	b	***SE***	***Z***	***OR***	***b***	***SE***	***Z***	***OR***	***b***	***SE***	***Z***	***OR***
TDL 1	−0.53	0.02	28.06	0.59	−0.59	0.02	29.12	0.55	−0.57	0.02	23.32	0.57
TDL 2	0.33	0.01	−24.25	1.39	0.34	0.01	−22.68	1.41	0.34	0.02	−21.26	1.41
TDL 3	1.92	0.02	−87.94	6.82	1.94	0.02	−86.20	6.96	1.92	0.03	−71.49	6.82
Shirking					−1.75	0.03	−55.33	0.17	−1.93	0.05	−37.74	0.15
Acting					0.46	0.02	29.42	1.58	0.56	0.03	17.69	1.75
TDL 2**Shirking*									−0.34	0.06	5.43	0.71
TDL 3**Shirking*									−0.02	0.11	0.14	0.98
TDL 2**Acting*									0.12	0.04	−3.32	1.13
TDL 3**Acting*									0.14	0.04	−3.27	1.15

Note: TDL = task difficulty level; TDL 2 or 3 indicates tasks locating difficulty level 2 or 3; *Shirking* and *Acting* were compared to the style of *Reflecting*. OR = Odds-ratio. All the estimated coefficients except for TDL 3**Shirking* were statistically significant at $\alpha = .001$ or $\alpha = .01$.

Table 7 shows a summary of the three explanatory item response models. The models were compared using the relative model fit indices of the Akaike Information Criterion (AIC; Akaike 1987) and Bayesian Information Criterion (BIC; Schwarz 1978). The model fit indices indicated that Model 2 had the best fit with the smallest AIC and BIC values. Since Models 0 and 1 were nested within each other, a direct comparison between the models was made using the likelihood ratio (LR) test. Given the significant improvement in model fit ($D = 5827$, $p < 0.001$) and a large reduction in residual variance (0.24) from Model 0 to Model 1, we could statistically infer participants' problem-solving styles explained their PSTRE performance. Similarly, the LR test between Model 1 and Model 2 was also significant ($D = 59.4$; $p < 0.001$). However, residual variance did not change from Model 1 to Model 2, indicating that the interaction effects included in Model 2 did not contribute to the model significantly. These results suggest that the advantageous effect of the *Acting* style and the disadvantageous impact of the *Shirking* style on PSTRE performance were consistent regardless of how difficult PSTRE tasks were.

Table 7. Overview of the estimated explanatory item response theory models.

Model	**Predictors**			**AIC**	**BIC**	**Variance**	**LR Test**		
	Task	**Person**	**Interaction**				***df***	**D**	**Comparison**
Model 0	TDL			161,860	161,959	0.42			
Model 1	TDL	PSS		156,037	156,156	0.18	2	5827 ***	with Model 0
Model 2	TDL	PSS	TDL * PSS	155,986	156,144	0.18	4	59.4 ***	with Model 1

*** $p < 0.001$. Note: TDL = Task difficulty level; PSS = Problem-solving style; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; D = Deviance; LR = Likelihood ratio.

4. Discussion

This study aimed to develop a novel understanding of what types of problem-solving styles humans exhibit in TRE using log file data and how the styles identified are associated with humans' performance in TRE. The results disclosed three types of problem-solving styles in TRE: *Acting*, *Reflecting*, and *Shirking*. We also found the superiority of the *Acting* style as well as the inferiority of the *Shirking* style for technology-based problem solving, irrespective of problem difficulties.

Our results contribute to the current literature in several ways. First, the presence of the *Acting* and *Reflecting* styles provides new evidence to support that learning modes are

associated with humans' dispositions to solve problems in TRE. We found that some participants prefer to be involved in operations and explorations with problem environments, while others prefer to observe rather than act in technology-based problem scenarios. These inclinations are aligned with participants' preference for action (i.e., *Acting*) or reflection (i.e., *Reflecting*) when they process information (Kolb and Kolb 2009; Richmond and Cummings 2005). This is likely because information processing is commonly involved in the problem-solving process (Reed and Vallacher 2020; van Gog et al. 2020). As Simon (1978) argued, the problem-solving process can be understood from an information-processing perspective. Thus, learning modes could serve as a stepping stone to understanding and profiling participants' dispositions towards problem solving in TRE.

Second, the *Shirking* style expands our knowledge of humans' dispositions towards problem solving in TRE. The participants adhering to the style of *Shirking* displayed a behavioral preference of scarcely pondering at the beginning of problem solving and barely exploring a problem scenario during the problem-solving process. Unlike the *Acting* and *Reflecting* styles, the *Shirking* style is a newly emergent style that describes participants' avoidance of planning and actions in problem solving in TRE (D'Zurilla and Chang 1995; Shoss et al. 2016). To construct a deeper understanding of the *Shirking* style, we examined the average response time of the three style groups and found that the *Shirking* style group spent less time (1.19 min) than those with the *Acting* style (2.95 min) or *Reflecting* style (2.51 min). However, the average response time was far longer than five seconds, which was used as a constant threshold for the minimum amount of time needed to validly respond to a task (e.g., Goldhammer et al. 2016; Wise and Kong 2005). In this respect, the *Shirking* style is different from disengaged test-taking behavior, though being disengaged is common in low-stakes assessments, such as the PIAAC 2012 (Goldhammer et al. 2016; Ulitzsch et al. 2021). Since various factors (e.g., cognition and personality) may impact how people respond to technology-based problems (Feist and Barron 2003), future studies should collect more data to explore what factors are associated with the presence of the three problem-solving styles in TRE.

Third, by comparing the three problem-solving styles, we are able to better understand the role of early planning and explorations in problem solving in TRE. Participants with an *Acting* style outperformed the other participants in problem solving in TRE, which confirms the assertion that actively initiating action may be a requisite for solving problems (Kolb and Fry 1975). When participants explore problem scenarios, including intuitive trial and error and stable routines within simulated computer platforms, they would gain the necessary information for problem solving, and thus enhance their chances of finding correct solutions (Liu et al. 2011). Eichmann et al. (2019) suspected that challenging tasks may require tryouts before meaningful planning. In this study, we found that participants with the *Reflecting* style were able to solve problems at difficulty levels 1 and 2, while those with the *Acting* style were able to solve more challenging problems, at all difficulty levels 1–3. This finding indicates that persistent trials play a more critical role than early planning in conducting difficult tasks. Further, in this study, the *Acting* style group differed from the *Reflecting* style group in the rescaled interaction frequency (0.73 higher) and planning duration (0.79 lower), indicating that high interaction frequency might make up for a short planning duration when participants solved technology-related problems, not vice versa.

We also noted some limitations of the present study. First, we did not explore participants excluded from this study due to outliers. Removed participants might take time to think or plan but finally skip an item. Furthermore, excluded participants might give up or abandon any explorations at the beginning of an item. These patterns barely reveal individuals' problem-solving styles in TRE, which have been defined as dispositions regarding how they are inclined to interact with surrounding technology environments in this study. However, their relationship to motivation when participants performed the low-stakes PSTRE assessment could be investigated in future studies. Second, it is actually not known how the time between participants' view of a task and their first interaction is actually used for planning. Eichmann et al. (2019) used the duration of the longest interval

between two successive interactions to define planning. However, Albert and Steinberg (2011) argued that individuals complete their initial planning phase before taking their first interaction with a task. Thus, additional work is needed to further explore the mapping of implicit planning processes. Third, we only abstracted planning duration and interaction frequency from log files corresponding to the *Acting* and *Reflecting* styles. Other learning styles described in ELT, such as Feeling and Thinking, were not included. Thus, this study partially confirms the applicability of ELT in describing problem-solving styles in TRE. Future research may include additionally detailed behavioral and/or cognitive information so that other styles and their potential link with PSTRE performance can be figured out. Fourth, this study only examined interaction effects between problem-solving styles and task difficulty levels on participants' performance, so future studies could include other critical cognitive factors, such as respondents' literacy and numeracy ability. As suggested by Xiao et al. (2019), cognitive factors may interact with participants' problem-solving styles and collectively act on individuals' problem-solving performance in TRE. Future studies could continue to explore potential interactions using the present research framework.

To summarize, this study provides critical evidence for the dominant role of active explorations in solving technology-based problems. The participants were adults so the knowledge generated in this study would help improve adult education programs, as well as computer-assisted problem-solving practice systems. As Ibieta et al. (2019) indicated, providing more detailed and specific cues (e.g., if you need to view emails, please click on this button) to facilitate participants' explorations and operations may be an effective approach in improving adults' problem-solving proficiency in TRE.

Author Contributions: Conceptualization, Y.G. and X.Z.; methodology, Y.G. and O.B.; software, Y.G. and O.B.; validation, X.Z., O.B. and Y.C.; formal analysis, Y.G. and X.S.; investigation, Y.C.; resources, Y.G.; data curation, Y.G.; writing—original draft preparation, Y.G.; writing—review and editing, X.Z. and Y.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: Ethical review and approval were waived for this study, due to an archival data set being used for the analyses.

Informed Consent Statement: Participant consent was waived due to the researchers only receiving a de-identified data set for their secondary analysis.

Data Availability Statement: The data presented in this study are not publicly available because they are confidential and proprietary (i.e., owned by the OECD). Requests to access the data should be directed to the OECD.

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

Appendix A

Table A1. An Exemplary Log File Data Including Events and Timestamps.

Item Name	Event Name	Event Type	Timestamp
U23x000S	taoPIAAC	START	0
U23x000S	taoPIAAC	NEXT_INQUIRY	14,449
U23x000S	taoPIAAC	NEXT_BUTTON	14,449
U23x000S	stimulus	CONFIRMATION_OPENED	14,452
U23x000S	taoPIAAC	BUTTON	14,454
U23x000S	taoPIAAC	DOACTION	14,454
U23x000S	stimulus	BUTTON	24,235
U23x000S	stimulus	CONFIRMATION_CLOSED	24,236
U23x000S	stimulus	DOACTION	24,236

Table A1. *Cont.*

Item Name	Event Name	Event Type	Timestamp
U23x000S	stimulus	MAIL_VIEWED	44,710
U23x000S	stimulus	MAIL_VIEWED	75,883
U23x000S	stimulus	MAIL_VIEWED	82,687
U23x000S	stimulus	MAIL_VIEWED	90,234
U23x000S	stimulus	MAIL_VIEWED	95,535
U23x000S	stimulus	MAIL_VIEWED	102,879
U23x000S	stimulus	MAIL_VIEWED	117,178
U23x000S	stimulus	MAIL_VIEWED	125,317
U23x000S	stimulus	MAIL_VIEWED	128,700
U23x000S	stimulus	FOLDER_VIEWED	141,563
U23x000S	stimulus	MAIL_DRAG	149,706
U23x000S	stimulus	MAIL_VIEWED	151,488
U23x000S	stimulus	TOOLBAR	165,881
U23x000S	stimulus	ENVIRONMENT	165,883
U23x000S	stimulus	DOACTION	165,883
U23x000S	stimulus	DOACTION	165,884
U23x000S	stimulus	DOACTION	165,884
U23x000S	stimulus	DOACTION	165,885
U23x000S	stimulus	TOOLBAR	167,934
U23x000S	stimulus	ENVIRONMENT	167,936
U23x000S	stimulus	DOACTION	167,936
U23x000S	stimulus	DOACTION	167,941
U23x000S	stimulus	DOACTION	167,942
U23x000S	stimulus	DOACTION	167,943
U23x000S	stimulus	TOOLBAR	171,676
U23x000S	stimulus	ENVIRONMENT	171,677
U23x000S	stimulus	DOACTION	171,677
U23x000S	stimulus	DOACTION	171,678
U23x000S	stimulus	DOACTION	171,679
U23x000S	stimulus	DOACTION	171,679
U23x000S	stimulus	TOOLBAR	173,631
U23x000S	stimulus	ENVIRONMENT	173,633
U23x000S	stimulus	DOACTION	173,633
U23x000S	stimulus	DOACTION	173,633
U23x000S	stimulus	DOACTION	173,634
U23x000S	stimulus	DOACTION	173,634
U23x000S	stimulus	TEXTLINK	182,570
U23x000S	stimulus	HISTORY_ADD	182,727
U23x000S	taoPIAAC	NEXT_INQUIRY	188,529
U23x000S	taoPIAAC	NEXT_BUTTON	188,529
U23x000S	stimulus	CONFIRMATION_OPENED	188,532
U23x000S	taoPIAAC	BUTTON	188,538
U23x000S	taoPIAAC	DOACTION	188,538
U23x000S	stimulus	BUTTON	190,901
U23x000S	stimulus	CONFIRMATION_CLOSED	190,902
U23x000S	taoPIAAC	NEXT_ITEM	190,904
U23x000S	taoPIAAC	END	190,905

Appendix B

Table A2. Difficulty Scores and Difficulty Levels of the 14 Tasks (Organisation for Economic Co-operation and Development (OECD) 2016).

Task	Difficulty Score	Difficulty Level	Difficulty Range	Average (*SD*)
1	286	1	268 to 286	274.0 (10.39)
10	286			
11	268			
2	299	2	296 to 325	311.7 (11.57)
4	316			
7	325			
8	305			
12	296			
13	320			
14	321			
3	346	3	342 to 374	354.2 (14.24)
5	374			
6	342			
9	355			

References

Akaike, Hirotugu. 1987. Factor analysis and AIC. *Psychometrika* 52: 317–32. [CrossRef]

Albert, Dustin, and Laurence Steinberg. 2011. Age differences in strategic planning as indexed by the Tower of London. *Child Development* 82: 1501–17. [CrossRef] [PubMed]

Bontchev, Boyan, Dessislava Vassileva, Adelina Aleksieva-Petrova, and Milen Petrov. 2018. Playing styles based on experiential learning theory. *Computers in Human Behavior* 85: 319–28. [CrossRef]

Botelho, Wagner Tanaka, Maria das Graças Bruno Marietto, João Carlos da Motta Ferreira, and Edson Pinheiro Pimentel. 2016. Kolb's experiential learning theory and Belhot's learning cycle guiding the use of computer simulation in engineering education: A pedagogical proposal to shift toward an experiential pedagogy. *Computer Applications in Engineering Education* 24: 79–88. [CrossRef]

Bulut, Okan. 2021. Eirm: Explanatory Item Response Modeling for Dichotomous and Polytomous Item Responses. R Package Version 0.4. Available online: https://CRAN.R-project.org/packages=eirm (accessed on 11 August 2021).

Bulut, Okan, Guher Gorgun, and Seyma Nur Yildirim-Erbasli. 2021. Estimating explanatory extensions of dichotomous and polytomous Rasch models: The eirm package in R. *Psych* 3: 308–21. [CrossRef]

Bunderson, C. Victor, Dillon K. Inouye, and James B. Olsen. 1989. The four generations of computerized educational measurement. In *Educational Measurement*. Edited by Robert L. Linn. New York: American Council on Education, Macmillan Publishing Co., pp. 367–407.

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software* 61: 1–36. [CrossRef]

D'Zurilla, Thomas J., and Edward C. Chang. 1995. The relations between social problem solving and coping. *Cognitive Therapy and Research* 19: 547–62. [CrossRef]

De Boeck, Paul, and Mark Wilson. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Statistics for Social Science and Public Policy. New York: Springer.

Eichmann, Beate, Frank Goldhammer, Samuel Greiff, Liene Pucite, and Johannes Naumann. 2019. The role of planning in complex problem solving. *Computers & Education* 128: 1–12. [CrossRef]

Eickmann, Paul, Alice Y. Kolb, and David A. Kolb. 2004. Designing learning. In *Managing as Designing: Creating a New Vocabulary for Management Education and Research*. Edited by Fred Collopy and Richird Boland. Stanford: Stanford University Press, pp. 241–47.

Feist, Gregory J., and Frank X. Barron. 2003. Predicting creativity from early to late adulthood: Intellect, potential, and personality. *Journal of Research in Personality* 37: 62–88. [CrossRef]

George, Darren, and Paul Mallery. 2010. *SPSS for Windows Step by Step: A Simple Guide and Reference*. Boston: Pearson.

Goldhammer, Frank, Thomas Martens, Gabriela Christoph, and Oliver Lüdtke. 2016. *Test-Taking Engagement in PIAAC (OECD Education Working Papers, No. 133)*. Paris: OECD Publishing.

Greiff, Samuel, Christoph Niepel, Ronny Scherer, and Romain Martin. 2016. Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior* 61: 36–46. [CrossRef]

Hämäläinen, Raija, Bram De Wever, Antero Malin, and Sebastiano Cincinnato. 2015. Education and working life: VET adults' problem-solving skills in technology-rich environments. *Computers & Education* 88: 38–47. [CrossRef]

Han, Zhuangzhuang, Qiwei He, and Matthias von Davier. 2019. Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests. *Frontiers in Psychology* 10: 2461. [CrossRef]

He, Qiwei, Dandan Liao, and Hong Jiao. 2019. Clustering behavioral patterns using process data in PIAAC problem-solving items. In *Theoretical and Practical Advances in Computer-Based Educational Measurement*. Edited by Bernard Veldkamp and Cor Sluijter. Cham: Springer, pp. 189–212.

He, Qiwei, Francesca Borgonovi, and Marco Paccagnella. 2021. Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education* 166: 104170. [CrossRef]

Henry, David B., Patrick H. Tolan, and Deborah Gorman-Smith. 2005. Cluster analysis in family psychology research. *Journal of Family Psychology* 19: 121–32. [CrossRef] [PubMed]

Hung, Yu Hsin, Ray I. Chang, and Chun Fu Lin. 2016. Hybrid learning style identification and developing adaptive problem-solving learning activities. *Computers in Human Behavior* 55: 552–61. [CrossRef]

Ibieta, Andrea, J. Enrique Hinostroza, and Christian Labbé. 2019. Improving students' information problem-solving skills on the Web through explicit instruction and the use of customized search software. *Journal of Research on Technology in Education* 51: 217–38. [CrossRef]

Ifenthaler, Dirk. 2012. Determining the effectiveness of prompts for self-regulated learning in problem-solving scenarios. *Educational Technology & Society* 15: 38–52.

Iñiguez-Berrozpe, Tatiana, and Ellen Boeren. 2020. Twenty-first century skills for all: Adults and problem solving in technology rich environments. *Technology, Knowledge and Learning* 25: 929–51. [CrossRef]

Isaksen, Scott G., Astrid H. Kaufmann, and Bjørn T. Bakken. 2016. An examination of the personality constructs underlying dimensions of creative problem-solving style. *Journal of Creative Behavior* 50: 268–81. [CrossRef]

Jyoti, Kiran, and Satyaveer Singh. 2011. Data clustering approach to industrial process monitoring, fault detection and isolation. *International Journal of Computer Applications* 17: 41–45. [CrossRef]

Kassambara, Alboukadel, and Fabian Mundt. 2020. *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses*, R Package Version 1.0.7; Available online: https://CRAN.R-project.org/package=factoextra (accessed on 11 August 2021).

Kaufman, Leonard, and Peter J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken: John Wiley and Sons.

Kim, Minchi C., and Michael J. Hannafin. 2011. Scaffolding problem solving in technology-enhanced learning environments (TELEs): Bridging research and theory with practice. *Computers & Education* 56: 403–17. [CrossRef]

Koć-Januchta, Marta M., Tim N. Höffler, Helmut Prechtl, and Detlev Leutner. 2020. Is too much help an obstacle? Effects of interactivity and cognitive style on learning with dynamic versus non-dynamic visualizations with narrative explanations. *Educational Technology Research and Development* 68: 2971–90. [CrossRef]

Kodinariya, Trupti M., and Prashant R. Makwana. 2013. Review on determining number of cluster in K-means clustering. *International Journal of Advance Research in Computer Science and Management Studies* 1: 90–95.

Koehler, Adrie A., Timothy J. Newby, and Peggy A. Ertmer. 2017. Examining the role of Web 2.0 tools in supporting problem solving during case-based instruction. *Journal of Research on Technology in Education* 49: 182–97. [CrossRef]

Koivisto, Jaana-Maija, Hannele Niemi, Jari Multisilta, and Elina Eriksson. 2017. Nursing students' experiential learning processes using an online 3D simulation game. *Education and Information Technologies* 22: 383–98. [CrossRef]

Kolb, Alice Y., and David A. Kolb. 2005a. Learning styles and learning spaces: Enhancing experiential learning in higher education. *Academy of Management Learning and Education* 4: 193–212. [CrossRef]

Kolb, Alice Y., and David A. Kolb. 2005b. *The Kolb Learning Style Inventory—Version 3.1. Technical Specifications*. Boston: Hay Resource Direct.

Kolb, Alice Y., and David A. Kolb. 2009. Experiential learning theory: A dynamic, holistic approach to management learning, education and development. In *Handbook of Management Learning, Education and Development*. Edited by Steven J. Armstrong and Cynthia Fukami. London: Sage Publications, pp. 42–68.

Kolb, David A. 2015. *Experiential Learning: Experience as the Source of Learning and Development*. Upper Saddle River: Pearson.

Kolb, David Allen, and Ronald Eugene Fry. 1975. *Toward an Applied Theory of Experiential Learning*. Cambridge: MIT Alfred P. Sloan School of Management.

Lewis, Tracy L., and Wanda J. Smith. 2008. Creating high performing software engineering teams: The impact of problem solving style dominance on group conflict and performance. *Journal of Computing Sciences in Colleges* 24: 121–29.

Liao, Dandan, Qiwei He, and Hong Jiao. 2019. Mapping background variables with sequential patterns in problem-solving environments: An investigation of United States adults' employment status in PIAAC. *Frontiers in Psychology* 10: 646. [CrossRef]

Liu, Chen-Chung, Yuan-Bang Cheng, and Chia-Wen Huang. 2011. The effect of simulation games on the learning of computational problem solving. *Computers & Education* 57: 1907–18. [CrossRef]

Millar, Roberto J., Shalini Sahoo, Takashi Yamashita, and Phyllis Cummins. 2020. Problem solving in technology-rich environments and self-rated health among adults in the U.S.: An analysis of the program for the international assessment of adult competencies. *Journal of Applied Gerontology* 39: 889–97. [CrossRef]

Morris, Thomas Howard. 2020. Experiential learning—A systematic review and revision of Kolb's model. *Interactive Learning Environments* 28: 1064–77. [CrossRef]

Nygren, Henrik, Kari Nissinen, Raija Hämäläinen, and Bram De Wever. 2019. Lifelong learning: Formal, non-formal and informal learning in the context of the use of problem-solving skills in technology-rich environments. *British Journal of Educational Technology* 50: 1759–70. [CrossRef]

Organisation for Economic Co-operation and Development (OECD). 2012. *Literacy, Numeracy, and Problem Solving in Technology-Rich Environments: Framework for the OECD Survey of Adult Skills*. Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (OECD). 2013. *OECD Skills Outlook 2013: First Results from the Survey of Adult Skills*. Paris: OECD Publishing, Available online: http://dx.doi.org/10.1787/9789264204256-en (accessed on 26 December 2021).

Organisation for Economic Co-operation and Development (OECD). 2016. *Skills Matter: Further Results from the Survey of Adult Skills*. Paris: OECD Publishing, Available online: http://dx.doi.org/10.1787/9789264258051-en (accessed on 26 December 2021).

Organisation for Economic Co-operation and Development (OECD). 2019. *Skills Matter: Additional Results from the Survey of Adult Skills*. Paris: OECD Publishing, Available online: https://doi.org/10.1787/1f029d8f-en (accessed on 26 December 2021).

Organisation for Economic Co-operation and Development (OECD). n.d. Job Search Part 1. Available online: https://piaac-logdata.tba-hosting.de/public/problemsolving/JobSearchPart1/pages/jsp1-home.html (accessed on 20 November 2021).

Oshima, Jun, and H. Ulrich Hoppe. 2021. Finding meaning in log-file data. In *International Handbook of Computer-Supported Collaborative Learning*. Edited by Ulrike Cress, Carolyn Rosé, Alyssa Friend Wise and Jun Oshima. Cham: Springer, pp. 569–84. [CrossRef]

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, Available online: https://www.R-project.org/ (accessed on 11 August 2021).

Reed, Stephen K., and Robin R. Vallacher. 2020. A comparison of information processing and dynamical systems perspectives on problem solving. *Thinking & Reasoning* 26: 254–90. [CrossRef]

Richmond, Aaron S., and Rhoda Cummings. 2005. Implementing Kolb's learning styles into online distance education. *International Journal of Technology in Teaching and Learning* 1: 45–54.

Romero, Jose Eulogio, Bennett J. Tepper, and Linda A. Tetrault. 1992. Development and validation of new scales to measure Kolb's 1985 learning style dimensions. *Educational and Psychological Measurement* 52: 171–80. [CrossRef]

Schwarz, Gideon. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–64. [CrossRef]

Selby, Edwin C., Donald J. Treffinger, Scott G. Isaksen, and Kenneth J. Lauer. 2004. Defining and assessing problem-solving style: Design and development of a new tool. *Journal of Creative Behavior* 38: 221–43. [CrossRef]

Sharma, Garima, and David A. Kolb. 2010. The learning flexibility index: Assessing contextual flexibility in learning style. In *Style Differences in Cognition, Learning, and Management: Theory, Research, and Practice*. Edited by Stephen Rayner and Eva Cools. London: Routledge, pp. 1–30. [CrossRef]

Shoss, Mindy K., Emily M. Hunter, and Lisa M. Penney. 2016. Avoiding the issue: Disengagement coping style and the personality-CWB link. *Human Performance* 29: 106–22. [CrossRef]

Simon, Herbert A. 1978. Information processing theory of human problem solving. In *Handbook of Learning and Cognitive Process*. Edited by D. Estes. Hillsdale: Lawrence Erlbaum Associates.

Tatnall, Arthur. 2014. ICT, education and older people in Australia: A socio-technical analysis. *Education and Information Technologies* 19: 549–64. [CrossRef]

Treffinger, Donald J., Edwin C. Selby, and Scott G. Isaksen. 2008. Understanding individual problem-solving style: A key to learning and applying creative problem solving. *Learning and Individual Differences* 18: 390–401. [CrossRef]

Ulitzsch, Esther, Qiwei He, and Steffi Pohl. 2021. Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics* 47: 3–35. [CrossRef]

van Gog, Tamara, Vincent Hoogerheide, and Milou van Harsel. 2020. The role of mental effort in fostering self-regulated learning with problem-solving tasks. *Educational Psychology Review* 32: 1055–72. [CrossRef]

Wang, Yingxu, and Vincent Chiew. 2010. On the cognitive process of human problem solving. *Cognitive Systems Research* 11: 81–92. [CrossRef]

Hadley Wickham, Romain Francois, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*, R Package Version 1.0.7; Available online: https://CRAN.R-project.org/package=dplyr (accessed on 11 August 2021).

Wise, Steven L., and Xiaojing Kong. 2005. Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education* 18: 163–83. [CrossRef]

Xiao, Feiya, Lucy Barnard-Brak, William Lan, and Hansel Burley. 2019. Examining problem-solving skills in technology-rich environments as related to numeracy and literacy. *International Journal of Lifelong Education* 38: 327–38. [CrossRef]

Zheng, Yingqin, Mathias Hatakka, Sundeep Sahay, and Annika Andersson. 2017. Conceptualizing development in information and communication technology for development (ICT4D). *Information Technology for Development* 24: 1–14. [CrossRef]

Zoanetti, Nathan, and Patrick Griffin. 2014. Log-file data as indicators for problem-solving processes. In *The Nature of Problem Solving*. Edited by Joachim Funke and Ben Csapo. Pairs: OECD.

Journal of
Intelligence

MDPI

Article

Shaky Student Growth? A Comparison of Robust Bayesian Learning Progress Estimation Methods

Boris Forthmann *, Natalie Förster and Elmar Souvignier

Institute of Psychology in Education, University of Münster, 48149 Münster, Germany; natalie.foerster@uni-muenster.de (N.F.); elmar.souvignier@uni-muenster.de (E.S.)
* Correspondence: boris.forthmann@wwu.de

Abstract: Monitoring the progress of student learning is an important part of teachers' data-based decision making. One such tool that can equip teachers with information about students' learning progress throughout the school year and thus facilitate monitoring and instructional decision making is learning progress assessments. In practical contexts and research, estimating learning progress has relied on approaches that seek to estimate progress either for each student separately or within overarching model frameworks, such as latent growth modeling. Two recently emerging lines of research for separately estimating student growth have examined robust estimation (to account for outliers) and Bayesian approaches (as opposed to commonly used frequentist methods). The aim of this work was to combine these approaches (i.e., robust Bayesian estimation) and extend these lines of research to the framework of linear latent growth models. In a sample of $N = 4970$ second-grade students who worked on the quop-L2 test battery (to assess reading comprehension) at eight measurement points, we compared three Bayesian linear latent growth models: (a) a Gaussian model, (b) a model based on Student's t-distribution (i.e., a robust model), and (c) an asymmetric Laplace model (i.e., Bayesian quantile regression and an alternative robust model). Based on leave-one-out cross-validation and posterior predictive model checking, we found that both robust models outperformed the Gaussian model, and both robust models performed comparably well. While the Student's t model performed statistically slightly better (yet not substantially so), the asymmetric Laplace model yielded somewhat more realistic posterior predictive samples and a higher degree of measurement precision (i.e., for those estimates that were either associated with the lowest or highest degree of measurement precision). The findings are discussed for the context of learning progress assessment.

Keywords: progress monitoring; Bayesian analysis; slope; growth; robust estimation

Citation: Forthmann, Boris, Natalie Förster, and Elmar Souvignier. 2022. Shaky Student Growth? A Comparison of Robust Bayesian Learning Progress Estimation Methods. *Journal of Intelligence* 10: 16. https://doi.org/10.3390/jintelligence10010016

Received: 31 October 2021
Accepted: 24 February 2022
Published: 1 March 2022

1. Introduction

The term progress monitoring refers to systematically gathering information on students' learning progress to guide feedback and instructional decision making. A prominent example of progress monitoring is curriculum-based measurement (CBM; Deno 1985), occurring in the context of special education. In CBM, parallel weekly assessments of core competencies such as reading are used to assess students' responsiveness to teachers' instructional decisions. An important feature of CBM is that assessments are indicators of and interpreted in relation to a desired learning goal (Fuchs 2004). Another similar form of progress monitoring is learning progress assessment (LPA), which refers to progress monitoring in everyday classrooms. LPA as implemented by the assessment system quop (Souvignier et al. 2021), for example, has longer time intervals between successive measurement points as compared to CBM. In addition, LPA tries to balance differentiated assessment of relevant skills (i.e., math and reading achievement) to allow differentiated feedback and acceptable psychometric properties. For example, the quop-L2 test series for reading assessment in second grade includes three subscales at all levels of language

(i.e., the word, sentence, and text levels; Förster et al. 2021; Förster and Kuhn 2021). If a student performs well at the word level but poorly at the sentence and text levels, the fit of instruction would be high if the teacher supports the student's sentence reading before supporting more complex higher-order reading comprehension strategies. The success of such progress monitoring implementations (regardless of whether CBM or LPA) can be evaluated via estimates of learning progress (Fuchs 2004).

1.1. Estimation of Learning Progress

The idea of estimating learning progress using the slope of a student's data plotted in a bivariate scatterplot can be traced back to work prior to the emergence of CBM (Deno and Mirkin 1977). Conceptually, the linear slope that occurs when plotting student performance against measurement time points allows one to assess the student's average learning progress over time (Silberglitt and Hintze 2007). While numerous methods exist for estimating slopes (Ardoin et al. 2013), in the context of progress monitoring, researchers have most often used ordinary least squares estimation. Historically, ordinary least squares can be understood as having replaced other methods such as quarter-intersect or split-middle (both methods require splitting the data into two halves to identify the median of each halve which builds the basis for drawing the slope; split-middle further requires that the same number of points is situated below and above the line) as the default in progress monitoring. Quarter-intersect or split-middle were considered more applicable in early years of CBM practice when computational power was not regularly available in school settings and growth estimates had to be calculated and drawn by hand (Ardoin et al. 2013). In addition, it was demonstrated in simulation studies that ordinary least squares estimates outperform estimates based on the medians of splitted data (Christ et al. 2012). Most recently, researchers have discussed and examined approaches that can be understood as either robust methods (e.g., non-parametric Theil–Sen regression; Bulut and Cormier 2018; Vannest et al. 2012) or Bayesian methods (Christ and Desjardins 2018; Solomon and Forsberg 2017). The ordinary least squares estimator makes assumptions (e.g., homoscedastic normally distributed errors) that can be violated in empirical data, and it is prone to influencing outliers. Indeed, the non-parametric Theil–Sen estimator does not require such strong assumptions and is robust with respect to outliers. Advantages of Bayesian estimation methods have been nicely summarized by Solomon and Forsberg (2017, p. 542): it can be robust, prior information can be utilized, and it has a natural compatibility with data-based decision making (i.e., posterior probabilities inform about intervention success).

Importantly, students may occasionally be tired or unmotivated when taking a test. In addition, researchers have identified that particular factors related to data collection (e.g., the place where the assessment is conducted, the person administering and/or scoring the test) may cause scores to fluctuate (Van Norman and Parker 2018). Hence, in the context of progress monitoring (i.e., repeated assessment of learning progress to inform feedback and instructional decision making), such fluctuations potentially influence performance at single measurement points and might yield single observations that strongly deviate from what might be expected (Bulut and Cormier 2018). Consequently, such outliers can influence estimates of student learning, especially when they occur at the beginning or toward the end of the period of assessment (Bulut and Cormier 2018). However, an accurate evaluation of student learning is critically important in the context of progress monitoring because such a data-based approach to decision making (Espin et al. 2017) relies on dynamic loops of assessment, instructional decisions, and feedback. To avoid this problem, Christ and Desjardins (2018) suggested using Bayesian slope estimation, which was found to be more precise and more realistic compared to ordinary least squares regression (Christ and Desjardins 2018).

1.2. Factors That Influence the Quality of Learning Progress Estimates

The quality of slope estimates in the context of progress monitoring does not only depend on the method of slope estimation. Both empirical and simulation studies have identified several other factors affecting the psychometric integrity of slope estimates such as measurement invariance, procedures of data collection, data collection schedules, and the number of measurement points. For a review of these factors from the perspective of CBM, see Ardoin et al. (2013).

Measurement invariance of the tests used is important to allow a straightforward interpretation of learning progress. While Ardoin et al. (2013) concluded that empirical tests of probe equivalence in CBMs are scarce in the literature, the importance of equivalent (i.e., parallel) tests has been emphasized in progress monitoring research. As recommended by Schurig et al. (2021, p. 2): " ... a good progress monitoring test should first check the dimensions, then the invariance ... ". The available evidence of CBM probes in terms of equivalence suggests that probes may not display form equivalence (Cummings et al. 2013) and findings indicated that psychometric quality of slope estimates depends on the chosen probe sets (Christ and Ardoin 2009). However, the quop-L2 test that was used in this work has demonstrated its factorial validity (Förster et al. 2021) and strong evidence in terms of practical equivalence based on a thorough item-response theory investigation focusing on accuracy and speed (Förster and Kuhn 2021), as well as strict measurement invariance when items are scored for efficiency of reading (Förster et al. 2021). The relevance of procedures of data collection has been already discussed in the introduction above. Clearly, variations in administration procedures can cause fluctuations in test performance, starting with varying times on the testing day at which tests are administered to a simple change of the testing room, for example. Beyond such potential influences on test performance, Bulut and Cormier (2018) thoroughly discussed progress monitoring schedules and the overall number of assessment points. They highlight that optimal schedules depend on the expected rate of improvement, which in turn can depends on various student characteristics. For example, from the perspective of CBM, a comprehensive simulation study revealed that validity and reliability of slope estimates depend on the overall duration (i.e., in weeks) of progress monitoring as well as the number of assessments within each week (Christ et al. 2013). Christ et al. (2013) found that valid and reliable slope estimation required at least four weeks of progress monitoring. While the overall duration of progress monitoring in LPA tends to be longer (e.g., 31 weeks in this study), the overall schedule must be considered to be clearly less dense with successive measurement timepoints being separated by approximately three-week intervals (Souvignier et al. 2021), for example. Beyond these aspects of the progress monitoring schedule, increasing the number of measurement points will increase the measurement precision of slope estimates. However, adding measurement timepoints close in time will, for most core skills, not result in huge information gains when it comes to slope assessment.

1.3. Aim of the Current Study

In this work, we aimed at employing robust Bayesian regression based on Student's *t*-distribution (Kruschke 2015) and Bayesian quantile regression (Yu and Moyeed 2001) to model the conditional median, as these approaches represent promising alternatives following both Christ and Desjardins's recommendation to use Bayesian estimation and Bulut and Cormier's call for robust slope estimation when outliers are present (Bulut and Cormier 2018). Importantly, estimating learning progress has relied on approaches that estimate progress for each student separately, but researchers have also applied overarching model frameworks, such as latent growth modeling, to progress monitoring data (Schatschneider et al. 2008; Yeo et al. 2012). Thus, the aim of this work was to combine robust and Bayesian learning progress estimation (i.e., robust Bayesian learning progress estimation) and extend these two recent lines of research toward the framework of linear latent growth models.

In this work, we explored the following research questions in the context of learning progress assessment of reading achievement in the second grade:

Research Question 1: Do robust Bayesian latent growth models outperform a simple Bayesian latent growth model based on the Gaussian distribution in terms of learning progress estimation?

Research Question 2: Which robust Bayesian latent growth model performs best in terms of learning progress estimation?

To answer these questions, we fitted Bayesian linear latent growth models based either on a Gaussian or a Student *t*-distribution to model reading comprehension efficiency. In addition, we added a Bayesian latent growth model based on an asymmetric Laplace model (i.e., Bayesian quantile regression) with the median as a conditional quantile to the set of candidate models.

2. Materials and Methods

2.1. Dataset

The dataset we used for this study was also used in a recent study on the reliability of learning progress estimates (Forthmann et al. 2021). In this dataset, N = 4970 second-grade students (age in years: M = 7.95, SD = 0.48; 53% boys and 47% girls) were assessed with the quop-L2 test series for reading achievement (Förster et al. 2021), quop-L2 comprising of equivalently constructed reading tests at all levels of language (i.e., the word, sentence, and text levels). Tests were administered at eight time points throughout the school year 2018/2019 via the assessment system quop (Souvignier et al. 2021). All tests were administered with about three-week intervals between successive tests starting in fall 2018, with two tests prior to the Christmas break. After Christmas break, four tests were administered prior to the Easter break in 2019 and the last two tests were administered between Easter break and Summer break. Thus, in total, the tests were completed over a period of 31 weeks. Initially, the cohort comprised of 6000 students. A total of 1030 students were excluded for the following reasons: (a) 140 students from international schools, (b) 227 students who were not in second grade but assigned to quop-L2, (c) three students who were younger than six years, (d) 94 students who were older than twelve years, (e) 333 students who had missing values on all measurement points, and (f) 233 identified duplicate cases.

For each item, subscale-specific quantiles were used as cut-offs for valid response behavior to correct for fast guessing (Wise 2017; Wise and DeMars 2010) and inacceptable slow responding. For fast guessing we used the 5%-quantile, whereas for slow responding we used the 99.5%-quantile. These quantiles were calculated for the complete sample comprising of all cohorts from 2015 to 2019 (N = 15,700) and for each subscale across all items (word level: lower bound = 1362.98 ms, upper bound = 41,032.86 ms; sentence level: lower bound = 1427.02 ms, upper bound = 53,742.18 ms; text level: lower bound = 877.36 ms, upper bound = 85,836.71 ms). These cut-offs were used prior to accuracy scoring of the items. In addition, the correct item summed residual time (CISRT) scoring was used as a measure of efficiency (Maris and van der Maas 2012). The time cut-offs were also useful for CISRT scoring which required item timing and quop-L2 tests are administered without any acute time limits. Item CISRT scores were averaged for each subscale (i.e., word, sentence, and text level) and scaled to be in the range from 0 to 10. Notably, a scoring of efficiency was used for two reasons: (a) it is in accordance with developmental models and empirical findings on reading skills, and (b) it allows assessing individual differences even among highly-proficient students. Developmental models of reading suggest that reading accuracy develops earlier, prior to automatizing the process into fluent reading (i.e., accurate and quick reading; Juul et al. 2014). The CISRT scoring mimics this by awarding fast reading only when the process was accurate. In addition, ceiling effects are likely for highly-proficient students in regular classes, when only accuracy is scored. Hence, introducing an additional speeded component or a scoring for efficiency for reading assessment is vital to measure individual differences in regular classrooms (Forthmann et al. 2020).

In this work, the more general construct of reading achievement was the focus (i.e., the higher order construct of reading at the word, sentence, and text level). Hence, scores at the word, sentence, and text level were used as observed indicators in a latent variable model to establish strong measurement invariance (Vandenberg and Lance 2000) prior to growth modeling. First, we established a configural model by comparing a simple longitudinal confirmatory factor analysis (CFA) model with one latent reading achievement variable based on the three observed scores at word, sentence, and text level at each of the eight timepoints. All latent covariances of the model were freely estimated, but residual covariances of the observed scores were fixed to a value of zero (Model 1). The loading of the word level indicator was fixed to a value of one at each timepoint to identify the model. We compared this configural model with another model that allowed residual covariances between scores at the same level. For example, all residual covariances between word-level scores were freely estimated, but cross-level covariances (e.g., between word-level and sentence-level scores) were fixed to zero (Model 2). As it turned out that residual covariances for sentence-level scores were mostly non-significant and rather small in size, we decided to add a more data driven and more parsimonious model that incorporated only residual covariances for word-level and text-level scores (Model 3). All models were estimated by the lavaan package (Rosseel 2012, version 0.6-9) for the statistical software R (R Core Team 2021, version 4.1.2 used on a local computer). We used full information maximum likelihood for model estimation to take missing values into account. This is justified by the fact that for this kind of longitudinal data, missing at random is the most likely underlying missing data mechanism (Asendorpf et al. 2014) and previous analyses of the missing data in quop-L2 also revealed patterns in accordance with missing at random (Förster et al. 2021). Robust maximum likelihood estimation was used to account for non-normality of the data. For general recommendations, with respect to model fit indices, we refer to relevant textbook chapters (West et al. 2012). For measurement invariance models we used the established criterion that the CFI should not decrease by more than .010 and complement change in CFI by change in RMSEA and SRMR (Chen 2007). Model 3 was chosen as the configural model based on highly comparable findings (Model 1: CFI = .920, RMSEA = .065, SRMR = .041; Model 2: CFI = .996, RMSEA = .019, SRMR = .010; Model 3: CFI = .994, RMSEA = .020, SRMR = .013). The decrease in CFI from this configural model to a strong invariance model (i.e., loadings and intercepts of the scores are constraint to be equal across time) was $-.006$, which was clearly smaller than the .010 criterion. In addition, this was accompanied by an increase of .007 for the RMSEA and .022 for the SRMR. We concluded that strong measurement invariance across time was quite reasonable for reading achievement measured by the subscales of quop-L2.

Next, we extracted factor scores by means of the Bartlett method (DiStefano et al. 2009) from the strong invariance model. To empirically justify the use of factor scores, we examined factor determinacy indices (FDI; Ferrando and Lorenzo-Seva 2018). FDIs were all excellent (all FDIs > .90) and allowed usage of factor scores even for individual assessments (Ferrando and Lorenzo-Seva 2018). These factor scores were the dependent variable (i.e., y_{pt}) in the latent growth models, as defined in Table 1. Hence, we have used a scoring based on standard maximum likelihood CFA in a first step and used Bayesian latent growth models in a second step. It should be mentioned that factor scores could be calculated for all students at all measurement timepoints. Hence, no missing values were present in the data when the latent growth models were estimated (see Section 2.2 below). We further estimated reliability by means of Cronbach's α (Cronbach 1951) and Bollen's ω_1 (Bollen 1980; Raykov 2001), as implemented in the semTools package (Jorgensen et al. 2021, version 0.5-5). Reliability estimates across measurement points were rather homogeneous (Cronbach's α: range from .74 to .78; Bollen's ω_1: range from .74 to .78). Hence, reliability of efficiency scores were clearly above the commonly cited .70 which is required for low-stakes decisions (Christ et al. 2005).

Table 1. Linear Latent Growth Model Definitions and Used Prior Distributions.

Model	Gaussian	Student's *t*	Asymmetric Laplace
Response Distribution	$y_{pt} \sim N(\eta_{pt}, \sigma^2)$	$y_{pt} \sim t(\eta_{pt}, \sigma^2, \nu)$	$y_{pt} \sim ALD(\eta_{pt}, \sigma^2, 0.50)$
Linear Predictor	$\eta_{pt} = \beta_{0,p} + \beta_{1,p}X_t$	$\eta_{pt} = \beta_{0,p} + \beta_{1,p}X_t$	$\eta_{pt} = \beta_{0,p} + \beta_{1,p}X_t$
Latent Variable Distribution	$\beta_p \sim MVN(\mu_\beta, \Sigma_\beta)$	$\beta_p \sim MVN(\mu_\beta, \Sigma_\beta)$	$\beta_p \sim MVN(\mu_\beta, \Sigma_\beta)$
Prior for μ_{β_0}	$t(0.3, 2.5, 3)$	$t(0.3, 2.5, 3)$	$t(0.3, 2.5, 3)$
Prior for μ_{β_1}	Improper flat prior	Improper flat prior	Improper flat prior
Prior for σ_{β_0}	$ht(0, 2.5, 3)$	$ht(0, 2.5, 3)$	$ht(0, 2.5, 3)$
Prior for σ_{β_1}	$ht(0, 2.5, 3)$	$ht(0, 2.5, 3)$	$ht(0, 2.5, 3)$
Prior for correlation matrices	$lkj(1)$	$lkj(1)$	$lkj(1)$
Prior for σ	$ht(0, 2.5, 3)$	$ht(0, 2.5, 3)$	$ht(0, 2.5, 3)$
Prior for ν	-	$\Gamma(2, 0.10)$	-

y_{pt} = reading efficiency factor score for person p at timepoint t. η_{pt} = linear predictor for person p at timepoint t. σ^2 = Residual variance. ν = degrees of freedom of Student's t-distribution. $\beta_{0,p}$ = Intercept of person p (i.e., initial level of reading efficiency). $\beta_{1,p}$ = Slope of person p (i.e., learning progress in reading efficiency). X_t = Coding variable of measurement timepoint t ($X_1 = 0, X_2 = 1, \ldots, X_8 = 7$). β_p = Matrix of latent variables $\beta_{0,p}$ and $\beta_{1,p}$. μ_β = Vector of latent variable means μ_{β_0} (i.e., the average intercept across all persons) and μ_{β_1} (i.e., the average slope across all persons). Σ_β = Covariance matrix of latent variables $\beta_{0,p}$ and $\beta_{1,p}$. $N()$ = Normal distribution. $t()$ = Student's t distribution. $ALD()$ = Asymmetric Laplace distribution. $MVN()$ = Multivariate normal distribution. $ht()$ = Half-t distribution. $lkj()$ = Lewandowski-Kurowicka-Joe distribution. $\Gamma()$ = Gamma distribution.

Finally, to make the comparisons presented in this work worthwhile, we examined if outliers were actually present in the data. Based on Mahalanobis distance (Tabachnick and Fidell 2005), we identified n = 229 students as multivariate outliers (i.e., approximately 5% of the total sample).

2.2. Analytical Approach

All models were fitted with the brms package (Bürkner 2017, 2018) for the statistical software R (R Core Team 2021). All models were estimated on the computer cluster of of the University of Münster (https://www.uni-muenster.de/IT/services/unterstuetzungsleistung/hpc/; accessed on 20 February 2022), and all distributions needed for the current research (Gaussian, Student's t, and asymmetric Laplace distributions) were implemented in brms. For each of the distributions, a linear growth model was specified in the brms model formula syntax. Specifically, the eight measurement points were coded using the numbers 0 to 7, which allowed for interpreting the intercept in the model as the initial value of reading efficiency. Hence, slope estimates represent the average progress after a three-week interval. Average reading progress between successive measurement points was represented in the model by the parameter μ_{β_1}. Intercept and slope were further modeled as latent variables to allow variation of initial level and learning progress across students. The correlation between intercepts and slopes was also estimated. The exact definitions and used priors for all three models are provided in Table 1.

After a first round of estimating the models, we found that the scale of the factor scores with an average close to zero was problematic for the estimation process. Specifically, we observed non-converging chains as flagged by $\hat{R}$ values around 1.50 and very, very low Bulk-ESS and Tail-ESS measures (all values were far below the recommended cut-offs; Vehtari et al. 2021). This was always observed for the asymmetric Laplace model and occasionally for the Student's t model. First, we experimented with increasing the number of iterations, but convergence issues as well as divergent transitions were still observed. Divergent transitions indicate that the MCMC algorithm cannot be trusted, and that the posterior distribution has not been well sampled (see here https://mc-stan.org/docs/2_19/reference-manual/divergent-transitions; accessed on 18 February 2022). Hence, we decided to simply multiply all factor scores by 30 (i.e., all factor scores at all timepoints) to update the fitted models (this value was arbitrarily increased after an initial attempt to multiply by ten, which indicated that scaling the values this way facilitated model estimation). In addition, we used 4000 iterations for the Gaussian growth model, whereas for the Student's t and asymmetric Laplace models, we needed 6000 iterations. We ran four

chains for each model. In addition, for the Student's *t* and asymmetric Laplace models, we had to set the control parameters max_treedepth and adapt_delta to values of 15 each and 0.95 or 0.90, respectively, to prevent divergent transitions.

Researchers have argued to pay close attention to various convergence diagnostics such as measures of potential scale reduction (PSR) and effective sample size to insure accurate Bayesian inference (e.g., Vehtari et al. 2021; Zitzmann and Hecht 2019). In our work, we used the improved $\hat{R}$ (i.e., PSR), Bulk-ESS, and Tail-ESS convergence statistics proposed and studied by Vehtari et al. (2021). These measures are implemented in Stan (Carpenter et al. 2017), on which brms (Bürkner 2017) is based and is immediately available in the model outputs. $\hat{R}$ should be <1.01 and all of our obtained $\hat{R}$ values were 1.00 (i.e., for all parameters in all models). In addition, Vehtari et al. (2021) recommend the ESS measures to be >400 when four chains are used (which we did for all models). The recommended value of 400 was surpassed for the Gaussian model (range of Bulk-ESS: 1296 to 4141; range of Tail-ESS: 1971 to 3470), the Student's *t* model (range of Bulk-ESS: 1497 to 9841; range of Tail-ESS: 2811 to 10,181), and the asymmetric Laplace model (range of Bulk-ESS: 1065 to 7590; range of Tail-ESS: 2264 to 8590).

Estimated model parameters were reported, along with 95% credible intervals. In brms, these intervals are based on the respective quantiles of the posterior samples. Cross-validation can be used for Bayesian multi-model inference (Sivula et al. 2020). Hence, models were compared based on approximate leave-one-out cross-validation (LOO; Vehtari et al. 2017). Approximate LOO was performed by Pareto smoothed importance sampling. We used the expected log-pointwise predictive density (ELPD; Vehtari et al. 2017) to evaluate the models' predictive accuracy. The ELPD difference and its standard error allow a profound evaluation of differences in terms of model fit (i.e., the difference can be interpreted in relation to the standard error).

Finally, to better understand differences between the models, graphical checks and correlational analyses were conducted. For all three models, we examined posterior predictive checking (Gelman et al. 2013). In addition, we looked at the densities of the Gaussian, Student's *t*, and asymmetric Laplace distributions based on the estimates obtained for the first measurement point of an average student; then, we checked the correlations between the estimates for the initial level and the learning progress based on the different models, and we compared estimates of measurement precision for the initial level and the learning progress between both robust approaches.

3. Results

3.1. Model Comparison and Model Parameter Findings

We found that the Student's *t* model performed best, as indicated by the LOO comparison results (i.e., as indicated by the value of zero for the ELPD difference; see Table 2). Notably, the ELPD difference between the Student's *t* and the asymmetric Laplace models was not larger in absolute size than twice its standard error (see Table 2), showing that both robust approaches performed equally well. The ELPD difference between the model based on Student's *t*-distribution and the simple Gaussian model was larger in absolute size than sixteen times its standard error, which represents very strong evidence in favor of the robust linear model. Given that the ELPD difference was immense when comparing the Student's *t* and the Gaussian models and negligible when comparing the Student's *t* and the asymmetric Laplace models, we concluded that both robust methods clearly outperform the Gaussian model.

In addition, the latent variable results were also more comparable when comparing both robust methods, and less similar for when each robust method was compared to the Gaussian model. At the same time, the overall pattern of findings was quite comparable. In terms of mean vectors, we found negative intercepts in all models (see Table 2), and the slope estimates were highly similar across all models. For the latent variable results (i.e., between person variation results), we found in all models that there were larger standard deviations for the intercepts than for the slopes. This hints at stronger interindividual dif-

ferences in the initial level of reading efficiency as compared to interindividual differences in average learning progress between successive measurement points. The correlation between random intercepts and slopes was negative and large in size (see Table 2). Hence, children with higher initial levels tended to make less learning progress over the school year in terms of reading efficiency. Finally, the models imply different levels of within-person variation as indicated by the residual variance estimates. Residual variance was highest for the Gaussian model and smallest for the Asymmetric Laplace model (see Table 2).

Table 2. Model Estimates and Comparisons for the Latent Growth Curve Models.

Model	Gaussian		Student's *t*		Asymmetric Laplace	
	Estimate	**95% CI**	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Person Level (Latent Variables)						
$\hat{\sigma}_{\beta_0}$	34.10	[33.32, 34.90]	34.28	[33.48, 35.06]	34.17	[33.38, 34.98]
$\hat{\sigma}_{\beta_1}$	2.79	[2.65, 2.94]	2.41	[2.28, 2.54]	2.29	[2.15, 2.42]
$\text{Cor}(\beta_{0,p}, \beta_{1,p})$	−0.55	[−0.58, −0.52]	−0.63	[−0.67, −0.60]	−0.63	[−0.66, −0.59]
Population Level						
μ_{β_0}	−13.98	[−15.00, −12.96]	−12.79	[−13.79, −11.75]	−12.70	[−13.74, −11.68]
μ_{β_1}	4.52	[4.40, 4.64]	4.59	[4.49, 4.70]	4.57	[4.46, 4.67]
σ	21.90	[21.73, 22.08]	14.55	[14.31, 14.80]	7.67	[7.59, 7.76]
ν	-	-	3.22	[3.08, 3.36]	-	-
Quantile	-	-	-	-	0.50	-
LOO Comparison						
ELPD Difference	−2600.60		0.00		−32.50	
ELPD Difference *SE*	154.00		0.00		30.40	

CI = credible interval. LOO = leave-one-out cross-validation. ELPD = expected log-pointwise predictive density. SE = standard error. Please see Table 1 for model definitions and equations.

3.2. Exploring Differences between the Models

We found that both robust modeling approaches performed similarly well in terms of the LOO comparison results (see Table 2). However, the Student's *t* and asymmetric Laplace distributions have different properties, implying differences between both models that deserve further exploration. For illustration and to facilitate a deeper understanding of the reported findings, in this section we also consider results for the Gaussian distribution.

Looking at graphical posterior predictive checking results (see Figure 1), it became apparent that both robust approaches were better able to model the peak around zero of the distribution of reading efficiency compared to the Gaussian model (see top row of plots in Figure 1). Upon visually inspecting the densities of observed (dark blue line) and sampled (light blue lines) reading efficiency within the range of observed values (the reading efficiency factor scores multiplied by 30 ranged from −286.11 to 135.88), we found no differences between both robust approaches in terms of behavior of predictive posterior samples (see top-middle and top-right plots in Figure 1).

Differences between both robust approaches only became visible when looking at the full ranges of predictive posterior samples (light blue boxes and dots), as depicted in the plots in the bottom row of Figure 1. Posterior predictive samples based on Student's *t*-distribution had ranges that well covered all outliers at the lower tail of the distribution of reading efficiency (for some draws, the lower-tail outliers were nicely replicated). However, the range of sampled values was clearly wider than the observed range of values; this was particularly the case for the upper tail of the distribution (see bottom-middle plot in Figure 1). Hence, the model based on Student's *t*-distribution produced outliers that strongly exceeded the most extreme cases of the observed data. Regarding the asymmetric Laplace model, this model was able to cover a portion of the extreme values at the lower tail of the distribution of reading efficiency factor scores (see bottom-right plot in Figure 1),

but the most extreme values were not fully covered. At the same time, posterior predictive samples based on the asymmetric Laplace distribution did not exceed the observed values as much as they did for Student's t at the upper tail. Hence, one might conclude that the asymmetric Laplace model yielded somewhat more realistic posterior predictive samples than the Student's t model did.

The Gaussian model (bottom-left plot in Figure 1) did not cover any of the most extreme values at the lower tail of the distribution of reading efficiency and—compared to both robust approaches—the box width of the original data was less well reproduced (this fact was harder to inspect by eye when looking at the density overlay plots in the top row of Figure 1). Consequently, posterior predictive checking provided further insights into the question of why both robust models performed better than the Gaussian model (i.e., they better reproduced the box width of the original data and better reproduced extreme values at the lower tail of the distribution) and how both robust approaches actually differed (i.e., the Student's t model better covered extreme values at the lower tail of the distribution, whereas the asymmetric Laplace model seemed to produce posterior predictive samples within the range of the observed values of reading efficiency).

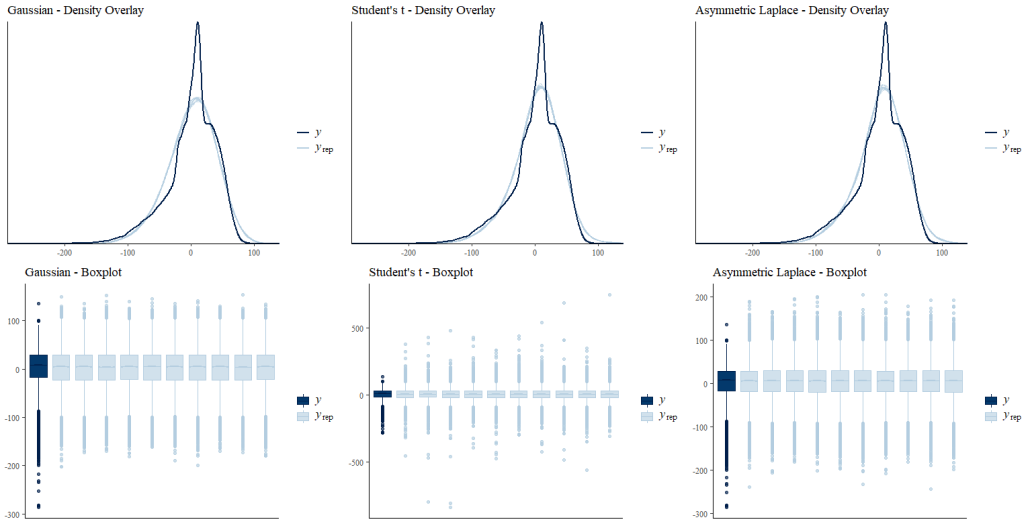

Figure 1. Graphical posterior predictive checking results. Top: Density overlay (based on ten posterior draws for each of the models) restricted to the range of −290 to 140 on the x axis to facilitate a comparison of model fit based on the main part of the empirical distribution (i.e., observed values of reading efficiency y ranged from −286.11 to 135.88). Bottom: Boxplots of the original data (dark blue) and ten draws of the posterior predictive distribution (boxes in light blue) to facilitate comparison of the sampled values between the three models.

We further examined the densities of the distributions, as based on the models reported in Table 2 for the first measurement point of an average student, to further understand how both robust models were able to better model the peak (and box width) of the original data distribution. The densities are depicted in Figure 2, where the left side of Figure 2 displays the full densities. Both robust models had more strongly peaked densities at the center of the distributions. In addition, the densities at the tails of the robust distributions had more probability mass than did those of the Gaussian distribution (see the right side of Figure 2). This explains exactly why these distributions work better when extreme values are present in the data: Such extreme values have a higher likelihood under these models and, hence, will have less influence on the model results as compared to the Gaussian model. This difference at the tails was clearly more pronounced for the Student's t model than the asymmetric Laplace model, which again suggests that the Student's t model was better able to handle the observed extreme values at the lower tail of the reading efficiency distribution (cf. Figure 1).

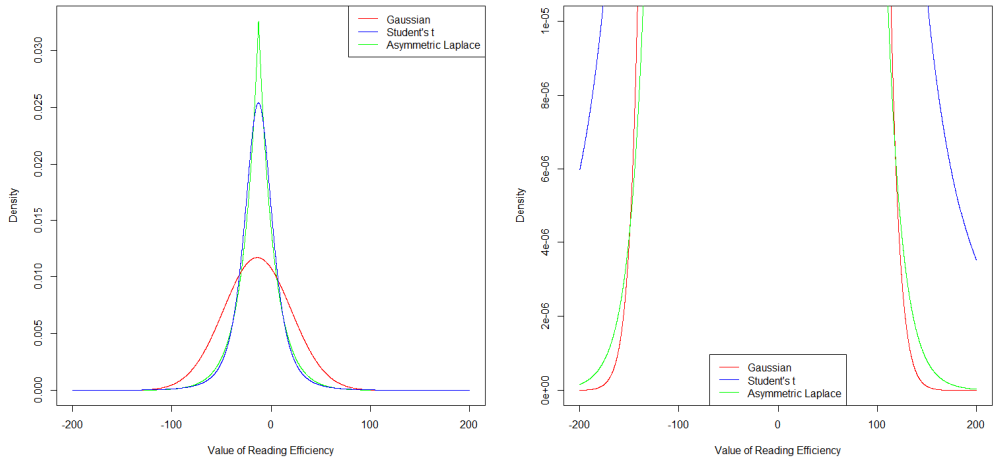

Figure 2. Distribution density plots for the following distributions (cf. Table 2): $N(-13.98, 34.10)$, t $(-12.79, 14.55, 3.22)$, and ALD $(-12.70, 7.67, 0.50)$. These are the distributions based on the estimated models reported in Table 2 for reading efficiency at the first measurement point for the average student. Left: y-axis range from 0 to 0.035 and x-axis range from -200 to 200. Right: "zoom-in" depiction of the densities to better visualize differences at the tails, i.e., y-axis range from 0 to 0.00001 and x-axis range from -200 to 200.

Next, we examined the correlations between the initial level (i.e., the random intercept estimates) and the learning progress estimates (i.e., the random slope estimates) between the three models (see Figure 3). These correlations allowed us to check whether the relative positioning of students with respect to important progress monitoring information differed between the models. The correlations between the initial level estimates are depicted in the top row of Figure 3; the initial level estimates based on both robust models correlated almost perfectly with the estimates obtained from the Gaussian model. In addition, initial level estimated based on both robust models revealed a perfect correlation (see top-right plot in Figure 3). Thus, estimating the initial level was quite robust across the three studied models. However, when looking at learning progress estimates, model choice clearly mattered for the relative positioning of students. The correlation between estimates based on the Gaussian model and both robust models was still large, but it was substantially lower ($r \approx .89$) than the correlation between learning progress estimates obtained from both robust models (the correlation was nearly perfect; see bottom-right plot in Figure 3). Thus, the differences between the Gaussian model and both robust models in terms of relative positioning were more strongly pronounced for learning progress than for initial level estimates, whereas both robust models yielded perfectly correlated estimates in this regard.

Finally, we investigated the measurement precision (i.e., the standard deviations of the posterior samples) of the initial level and the learning progress estimates based on both robust models. As both robust models make different distributional assumptions, the estimates of measurement precision for the initial level and the learning progress depended on the characteristics of these distributions (e.g., the asymmetric Laplace distribution was more peaked, whereas Student's t-distribution had heavier tails; see Figure 2 above). To compare measurement precision between both robust models, we divided the standard deviations of the posterior samples of initial level and learning progress estimates by their respectively estimated standard deviations of the respective latent variable distributions reported in Table 2. As depicted in Figure 4, we found that the measurement precision of the initial level estimates was clearly higher than the measurement precision of the learning progress estimates (compare the left and right plots in Figure 4). In addition, estimates of the measurement precision for both models had a positive yet non-linear relationship (see the LOESS curves depicted in red in both plots in Figure 4). At the lower and upper tails of measurement precision estimates, we observed that the measurement precision based on the asymmetric Laplace model was higher. Hence, in situations that allow for

choosing an asymmetric Laplace latent growth model, learning progress parameters for individual students (i.e., those estimates that tend to be associated with the lowest or highest measurement precision) would be expected to have greater measurement precision than when applying Student's *t*-distribution.

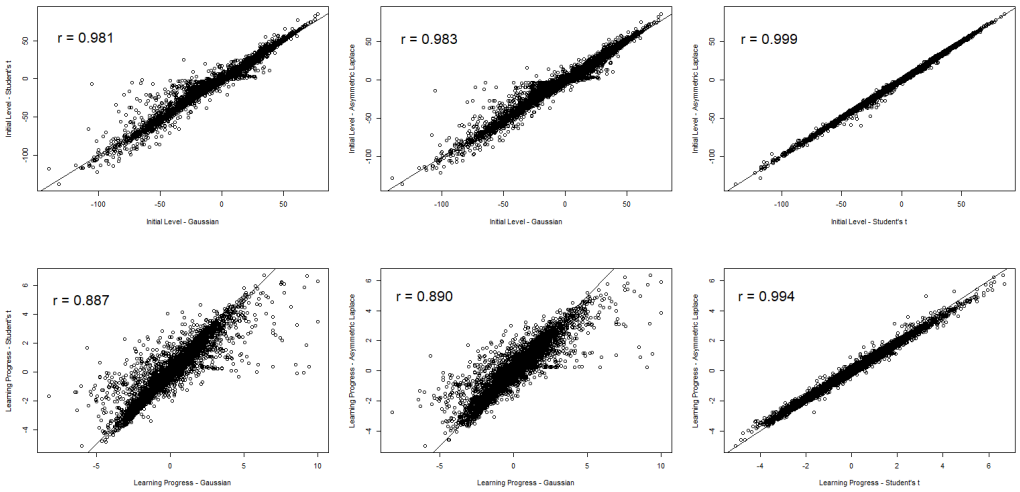

Figure 3. Bivariate scatter plots between initial level and learning progress estimates.

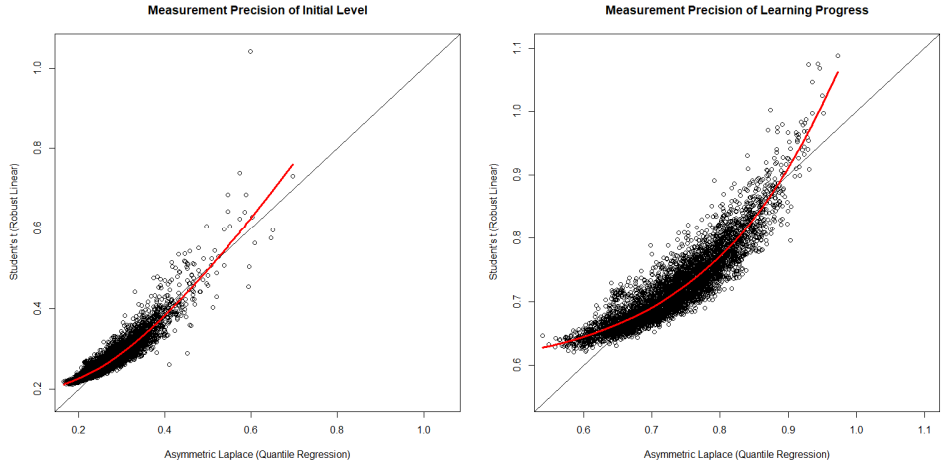

Figure 4. Bivariate scatter plots of measurement precision estimates. Red line = LOESS curve.

4. Discussion

In this work, we examined three Bayesian models to estimate student learning progress to extend recent calls for using Bayesian and robust approaches (Bulut and Cormier 2018; Christ and Desjardins 2018; Solomon and Forsberg 2017). We found that both robust models had better model-data fit than a Gaussian linear growth model did. Christ and Desjardins (2018) found that Bayesian Gaussian linear regression worked better than standard ordinary least squares regression for estimating the learning progress of individual students. Here, however, latent growth models were estimated and compared with each other. We found that linear growth models based on the Student's *t*-distribution and the asymmetric Laplace distribution performed very similarly in terms of LOO comparison and also in terms of relative positioning of students with respect to the initial level and the learning progress estimates, as revealed by nearly perfect correlations between the estimates of these two models. In addition, we showed that both robust models were better than the Gaussian model at modeling the peak and lower tail of the distribution of reading efficiency. Specifically, the likelihood of extreme values at the lower end of the tail of the distribution was higher for both robust models, meaning that these values are

not as influential on model estimation as they are for the Gaussian model (Kruschke 2015). Indeed, this behavior was more pronounced for the Student's *t* model. However, differences between these two models were also revealed. While the Student's *t* model performed statistically slightly better (in terms of LOO comparison and posterior predictive samples that well covered extreme values at the lower tail of the reading efficiency distribution), the asymmetric Laplace model yielded somewhat more realistic posterior predictive samples and gave a higher degree of measurement precision for those estimates that were associated with either the lowest or highest degrees of measurement precision.

These findings may generalize to other core competencies (e.g., math skills; Salaschek and Souvignier 2014; Boorse and Van Norman 2021), grade levels (e.g., reading comprehension in fourth grade; Förster and Souvignier 2014), and contexts (e.g., special needs educational setting; Jenkins et al. 2009) as long as the score distribution shows signs of being more peaked or having more extreme values at the tails of the distribution of progress monitoring data. However, we recommend that all three models be carefully examined in terms of model fit because the models have different distributional characteristics that imply, for example, differences in the measurement precision of learning progress monitoring estimates. Hence, the choice of model should be guided empirically, not based on favorable model properties per se (i.e., without any evaluation of model fit). Another option with this particular set of candidate models would be to test whether the conclusions drawn from analyses are robust across the models.

Despite differences between CBM and LPA (see Section 1), there are common aspects also. The estimation of learning progress (i.e., growth) by means of slopes, for example, is a technical feature that is shared by CBM and LPA. We are well aware that the difference in schedules has implications for the quality of slope estimates. However, we are confident that our study has implications even for CBM growth modeling, despite these differences. First, while CBM research on slope estimates often focuses on single-case data (e.g., Christ and Desjardins 2018; Solomon and Forsberg 2017), there are also latent growth model applications in the CBM literature (Keller-Margulis and Mercer 2014; Yeo et al. 2012). Importantly, for these applications, even larger intervals between successive measurements were allowed (e.g., testing in fall, winter, and spring) as compared to the ones used in our sample (approximately three-week intervals). Hence, given that CBM displays quite a range in terms of intervals between successive measurement points, our work has implications for such latent growth model applications within the CBM framework (i.e., at least when we consider potential differences in terms of inter-test intervals).

Beyond these potential differences (and similarities) between CBM and LPA applications of progress monitoring, however, we would like to highlight that the proof-of-concept provided by the empirical findings in our work are most likely to generalize when outliers are present in progress monitoring data. In our dataset we found that approximately 5% of the students qualified as multivariate outliers (this was further evident in the visual inspections of the data; see Figure 1). It is hard to say if this percentage of extreme cases is representative for progress monitoring data in general. However, for datasets with a larger number of extreme cases, it is clear that the current findings matter (perhaps even more). In addition, even for less influential cases, our work has clear implications. That can be easily seen when looking at slope estimation of single-case data that can be heavily influenced by only one influencing outlier at one of the measurement points (Bulut and Cormier 2018). Using a model based on Student's *t* distribution instead of a Gaussian model will be a much better choice in that situation as compared to a Gaussian model. In relation to this, Christ and Desjardins (2018) have examined and discussed the role of prior choice and concluded that choosing reasonable priors is crucial for Bayesian slope estimation to have an advantage in terms of measurement precision and realistic estimates over ordinary leas squares regression. Choosing reasonable priors might only result when knowledge about the distributions of both intercept and slope is available to inform Bayesian slope estimation of single-case data. Specifically, knowledge that is obtained from latent growth models applied to a large progress monitoring dataset, such as the one studied in this work,

seems to be particularly useful to construct reasonable priors for single-case estimation. Otherwise, it has been shown that such type of knowledge is not always needed to construct advantageous priors (Finch and Miller 2019; Zitzmann et al. 2021). For instance, Zitzmann et al. (2021) showed analytically that priors for variability parameters can perform well even when they are incorrect and do not represent the "true" parameter. Future research is needed to investigate such phenomena for robust Bayesian growth models in the context of progress monitoring.

This research was limited to reading comprehension in second-grade regular classrooms. In addition, we operationalized reading comprehension as a higher-order construct, but the quop-L2 test series also allows for a more detailed look at reading comprehension at the word, sentence, and text levels. Given that these subdimensions of quop-L2 might have different empirical distributions, it is not clear whether the findings generalize to each of the test's subdimensions. It should further be mentioned that we considered only linear growth models, as they fit nicely with the traditional conception of learning progress in the CBM literature (Silberglitt and Hintze 2007). However, researchers have also examined more complex growth models such as quadratic growth (Schatschneider et al. 2008). Hence, beyond the distribution of the used growth model, we consider it a promising path for future research to further compare different functional forms of the growth trajectory.

Notably, we used a scoring based on standard maximum likelihood CFA in a first step and used Bayesian latent growth models in a second step. While this might appear overly pragmatic in the sense that we did not hesitate to cross two different statistical philosophies within one and the same analysis, we are convinced that there are good reasons to believe that choosing this two-step approach has not undermined our aims. First, this initial step was undertaken to control for potential differences with respect to psychometric properties across timepoints (e.g., differences in intercepts of word, sentence, and text level scores could have influenced learning progress estimates within the latent growth curve model). The alternative would have been to use sum scores of observed scores, for example, which cannot be assumed to be of the same psychometric quality. Second, approximately 5% of the participants were identified as multivariate outliers which emphasizes the general need for robust modeling. Third, in the tradition of progress monitoring research, outliers are considered at the level of scores at each of the measurement timepoints (e.g., Bulut and Cormier 2018). The outliers in our work were also considered and taken into account by robust Bayesian latent growth models at the level of scores at different timepoints. One-step approaches in which latent variables are modeled at each measurement point and growth is modeled by means of higher-order latent variables would have shifted the question of outliers to the level of indicators of reading efficiency at each timepoint (one could even think of models in which the focus is shifted to item-level outliers). We argue that such a shift of the focus of outlier treatment would be worth future investigations, but it was beyond the aims of our current investigation.

Moreover, it should be discussed that we needed to fit all models on a high-performance computer cluster. A simple desktop or laptop would not have succeeded in this task within the same amount of time. Hence, while Bayesian model syntax building and model specification is clearly facilitated by the user-friendly implementation of brms (Bürkner 2017), special knowledge to run these models on a computer cluster is needed to make this workable for a large progress monitoring dataset, as studied in this work.

To conclude, the current work adds to the literature on learning progress estimation by combining the ideas of robust and Bayesian estimation into an overarching latent growth modeling framework. Here we showed that robust latent growth models outperform standard Gaussian models, and we found that these models were better capable of modeling a stronger peaked distribution and more extreme values at the lower tail of the distribution. We further found that the asymmetric Laplace model had, for some estimates, a higher degree of measurement precision, and yielded more realistic posterior predictive samples. These findings look promising for future applications, and we hope

that the outlined methods in this work will be tested and extended in the field of progress monitoring research.

Author Contributions: Conceptualization, B.F., N.F. and E.S.; Data curation, B.F.; Formal analysis, B.F.; Investigation, N.F.; Methodology, B.F.; Resources, E.S.; Software, B.F.; Visualization, B.F.; Writing—original draft, B.F.; Writing—review & editing, N.F. and E.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: This study was carried out in accordance with the recommendations by the ethics committee of the department of psychology of the University in Münster. An ethics approval was not required as per institutional and national guidelines.

Informed Consent Statement: For participants involved in this study, either informed consent was obtained from their parents or their participation was regulated based on a contractual regulation that allowed us to use participant data in an anonymized form for scientific purposes.

Data Availability Statement: The analyses scripts and data used for this study are openly available in the Open Science Framework: https://osf.io/hjx43/ (accessed on 18 February 2022).

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

References

Ardoin, Scott P., Theodore J. Christ, Laura S. Morena, Damien C. Cormier, and David A. Klingbeil. 2013. A systematic review and summarization of the recommendations and research surrounding Curriculum-Based Measurement of oral reading fluency (CBM-R) decision rules. *Journal of School Psychology* 51: 1–18. [CrossRef]

Asendorpf, Jens B., Rens van de Schoot, Jaap J. A. Denissen, and Roos Hutteman. 2014. Reducing bias due to systematic attrition in longitudinal studies: The benefits of multiple imputation. *International Journal of Behavioral Development* 38: 453–60. [CrossRef]

Bollen, Kenneth A. 1980. Issues in the Comparative Measurement of Political Democracy. *American Sociological Review* 45: 370. [CrossRef]

Boorse, Jaclin, and Ethan R. Van Norman. 2021. Modeling within-year growth on the Mathematics Measure of Academic Progress. *Psychology in the Schools* 58: 2255–68. [CrossRef]

Bulut, Okan, and Damien C. Cormier. 2018. Validity Evidence for Progress Monitoring With Star Reading: Slope Estimates, Administration Frequency, and Number of Data Points. *Frontiers in Education* 3. [CrossRef]

Bürkner, Paul-Christian. 2017. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80: 1–28. [CrossRef]

Bürkner, Paul-Christian. 2018. Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal* 10: 395. [CrossRef]

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Miachael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76: 1–32. [CrossRef]

Chen, Fang F. 2007. Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal* 14: 464–504. [CrossRef]

Christ, Theodore J., and Christopher D. Desjardins. 2018. Curriculum-Based Measurement of Reading: An Evaluation of Frequentist and Bayesian Methods to Model Progress Monitoring Data. *Journal of Psychoeducational Assessment* 36: 55–73. [CrossRef]

Christ, Theodore J., and Scott P. Ardoin. 2009. Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology* 47: 55–75. [CrossRef]

Christ, Theodore J., Cengiz Zopluoglu, Barbara D. Monaghen, and Ethan R. Van Norman. 2013. Curriculum-Based Measurement of Oral Reading: Multi-study evaluation of schedule, duration, and dataset quality on progress monitoring outcomes. *Journal of School Psychology* 51: 19–57. [CrossRef]

Christ, Theodore J., Cengiz Zopluoglu, Jeffery D. Long, and Barbara D. Monaghen. 2012. Curriculum-Based Measurement of Oral Reading: Quality of Progress Monitoring Outcomes. *Exceptional Children* 78: 356–73. [CrossRef]

Christ, Theodore J., Kristin N. Johnson-Gros, and John M. Hintze. 2005. An examination of alternate assessment durations when assessing multiple-skill computational fluency: The generalizability and dependability of curriculum-based outcomes within the context of educational decisions. *Psychology in the Schools* 42: 615–22. [CrossRef]

Cronbach, Lee J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334. [CrossRef]

Cummings, Kelli D., Yonghan Park, and Holle A. Bauer Schaper. 2013. Form Effects on DIBELS Next Oral Reading Fluency Progress-Monitoring Passages. *Assessment for Effective Intervention* 38: 91–104. [CrossRef]

Deno, Stanley L. 1985. Curriculum-Based Measurement: The Emerging Alternative. *Exceptional Children* 52: 219–32. [CrossRef] [PubMed]

Deno, Stanley L., and Phyllis K. Mirkin. 1977. *Data-Based Program Modification: A Manual*. Minneapolis: Leadership Training Institute for Special Education.

DiStefano, Christine, Min Zhu, and Diana Mîndrilã. 2009. Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Assessment, Research, and Evaluation* 14: 20. [CrossRef]

Espin, Christine A., Miya M. Wayman, Stanley L. Deno, Kristen L. McMaster, and Mark de Rooij. 2017. Data-Based Decision-Making: Developing a Method for Capturing Teachers' Understanding of CBM Graphs. *Learning Disabilities Research & Practice* 32: 8–21. [CrossRef]

Ferrando, Pere J., and Urbano Lorenzo-Seva. 2018. Assessing the Quality and Appropriateness of Factor Solutions and Factor Score Estimates in Exploratory Item Factor Analysis. *Educational and Psychological Measurement* 78: 762–80. [CrossRef]

Finch, W. Holmes, and J. E. Miller. 2019. The Use of Incorrect Informative Priors in the Estimation of MIMIC Model Parameters with Small Sample Sizes. *Structural Equation Modeling: A Multidisciplinary Journal* 26: 497–508. [CrossRef]

Förster, Natalie, and Elmar Souvignier. 2014. Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept. *Learning and Instruction* 32: 91–100. [CrossRef]

Förster, Natalie, and Jörg-Tobias Kuhn. 2021. Ice is hot and water is dry: Developing equivalent reading tests using rule-based item design. *European Journal of Psychological Assessment*. [CrossRef]

Förster, Natalie, Mathis Erichsen, and Boris Forthmann. 2021. Measuring Reading Progress in Second Grade: Psychometric Properties of the quop-L2 Test Series. *European Journal of Psychological Assessment*. [CrossRef]

Forthmann, Boris, Natalie Förster, and Elmar Souvignier. 2021. Empirical Reliability: A Simple-to-Calculate Alternative for Reliability Estimation of Growth in Progress Monitoring. *Manuscript Submitted for Publication*.

Forthmann, Boris, Rüdiger Grotjahn, Philipp Doebler, and Purya Baghaei. 2020. A Comparison of Different Item Response Theory Models for Scaling Speeded C-Tests. *Journal of Psychoeducational Assessment* 38: 692–705. [CrossRef]

Fuchs, Lynn S. 2004. The Past, Present, and Future of Curriculum-Based Measurement Research. *School Psychology Review* 33: 188–92. [CrossRef]

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*, 3rd ed. New York: Chapman and Hall/CRC.

Jenkins, Joseph R., J. Jason Graff, and Diana L. Miglioretti. 2009. Estimating Reading Growth Using Intermittent CBM Progress Monitoring. *Exceptional Children* 75: 151–63. [CrossRef]

Jorgensen, Terrence D., Sunthud Pornprasertmanit, Alexander M. Schoemann, and Yves Rosseel. 2021. semTools: Useful Tools for Structural Equation Modeling (R Package Version 0.5-4). Available online: https://cran.r-project.org/package=semTools (accessed on 18 February 2022).

Juul, Holger, Mads Poulsen, and Carsten Elbro. 2014. Separating speed from accuracy in beginning reading development. *Journal of Educational Psychology* 106: 1096–106. [CrossRef]

Keller-Margulis, Milena A., and Sterett H. Mercer. 2014. R-CBM in spanish and in english: Differential relations depending on student reading performance. *Psychology in the Schools* 51: 677–92. [CrossRef]

Kruschke, John K. 2015. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*, 2nd ed. New York: Academic Press.

Maris, Gunter, and Han van der Maas. 2012. Speed-Accuracy Response Models: Scoring Rules based on Response Time and Accuracy. *Psychometrika* 77: 615–33. [CrossRef]

Raykov, Tenko. 2001. Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology* 54: 315–23. [CrossRef]

R Core Team. 2021. *R: A Language and Environment for Statistical Computing (4.1.2)*. Vienna: R Foundation for Statistical Computing. Available online: https://www.r-project.org/ (accessed on 18 February 2022).

Rosseel, Yves. 2012. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software* 48: 2. [CrossRef]

Salaschek, Martin, and Eelmar Souvignier. 2014. Web-Based Mathematics Progress Monitoring in Second Grade. *Journal of Psychoeducational Assessment* 32: 710–24. [CrossRef]

Schatschneider, Christopher, Richard K. Wagner, and Elizabetz C. Crawford. 2008. The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences* 18: 308–15. [CrossRef]

Schurig, Michael, Jana Jungjohann, and Markus Gebhardt. 2021. Minimization of a Short Computer-Based Test in Reading. *Frontiers in Education* 6: 684595. [CrossRef]

Silberglitt, Benjamin, and John M. Hintze. 2007. How Much Growth Can We Expect? A Conditional Analysis of R—CBM Growth Rates by Level of Performance. *Exceptional Children* 74: 71–84. [CrossRef]

Sivula, Tuomas, Måns Magnusson, and Aki Vehtari. 2020. Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison. *arXiv* arXiv:2008.10296.

Solomon, Benjamin G., and Ole J. Forsberg. 2017. Bayesian asymmetric regression as a means to estimate and evaluate oral reading fluency slopes. *School Psychology Quarterly* 32: 539–51. [CrossRef]

Souvignier, Elmar, Natalie Förster, Karin Hebbecker, and Birgit Schütze. 2021. Using digital data to support teaching practice—Quop: An effective web-based approach to monitor student learning progress in reading and mathematics in entire classrooms. In *International Perspectives on School Settings, Education Policy and Digital Strategies. A Transatlantic Discourse in Education Research*. Edited by Sieglinde Jornitz and Annika Wilmers. Opladen: Budrich, pp. 283–98.

Tabachnick, Barbara G., and Linda S. Fidell. 2005. *Using Multivariate Statistics*, 5th ed. Boston: Pearson/Allyn and Bacon.

Van Norman, Ethan R., and David C. Parker. 2018. A Comparison of Split-Half and Multilevel Methods to Assess the Reliability of Progress Monitoring Outcomes. *Journal of Psychoeducational Assessment* 36: 616–27. [CrossRef]

Vandenberg, Robert J., and Charles E. Lance. 2000. A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods* 3: 4–70. [CrossRef]

Vannest, Kimberly J., Richard I. Parker, John L. Davis, Denise A. Soares, and Stacey L. Smith. 2012. The Theil–Sen Slope for High-Stakes Decisions from Progress Monitoring. *Behavioral Disorders* 37: 271–80. [CrossRef]

Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2021. Rank-Normalization, Folding, and Localization: An Improved Rˆ for Assessing Convergence of MCMC (with Discussion). *Bayesian Analysis* 16: 2. [CrossRef]

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27: 1413–32. [CrossRef]

West, Stephen G., Aaron B. Taylor, and Wei Wu. 2012. Model fit and model selection in structural equation modeling. In *Handbook of Structural Equation Modeling*. Edited by Rick H. Hoyle. New York: The Guilford Press, pp. 209–31.

Wise, Steven L. 2017. Rapid-Guessing Behavior: Its Identification, Interpretation, and Implications. *Educational Measurement: Issues and Practice* 36: 52–61. [CrossRef]

Wise, Steven L., and Christine E. DeMars. 2010. Examinee Noneffort and the Validity of Program Assessment Results. *Educational Assessment* 15: 27–41. [CrossRef]

Yeo, Seungsoo, Jamie Y. Fearrington, and Theodore J. Christ. 2012. Relation Between CBM-R and CBM-mR Slopes. *Assessment for Effective Intervention* 37: 147–58. [CrossRef]

Yu, Keming, and Rana A. Moyeed. 2001. Bayesian quantile regression. *Statistics & Probability Letters* 54: 437–47. [CrossRef]

Zitzmann, Steffen, and Martin Hecht. 2019. Going Beyond Convergence in Bayesian Estimation: Why Precision Matters Too and How to Assess It. *Structural Equation Modeling: A Multidisciplinary Journal* 26: 646–61. [CrossRef]

Zitzmann, Steffen, Oliver Lüdtke, Alexander Robitzsch, and Martin Hecht. 2021. On the Performance of Bayesian Approaches in Small Samples: A Comment on Smid, McNeish, Miocevic, and van de Schoot. *Structural Equation Modeling: A Multidisciplinary Journal* 28: 40–50. [CrossRef]

Journal of *Intelligence*

MDPI

Article

Estimating Local Structural Equation Models

Alexander Robitzsch [1,2]

[1] IPN–Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany; robitzsch@ipn.uni-kiel.de
[2] Centre for International Student Assessment (ZIB), Olshausenstraße 62, 24118 Kiel, Germany

Abstract: Local structural equation models (LSEM) are structural equation models that study model parameters as a function of a moderator. This article reviews and extends LSEM estimation methods and discusses the implementation in the R package sirt. In previous studies, LSEM was fitted as a sequence of models separately evaluated as each value of the moderator variables. In this article, a joint estimation approach is proposed that is a simultaneous estimation method across all moderator values and also allows some model parameters to be invariant with respect to the moderator. Moreover, sufficient details on the main estimation functions in the R package sirt are provided. The practical implementation of LSEM is demonstrated using illustrative datasets and an empirical example. Moreover, two simulation studies investigate the statistical properties of parameter estimation and significance testing in LSEM.

Keywords: local structural equation modeling; confirmatory factor analysis; differentiation; dedifferentiation; invariance

Citation: Robitzsch, Alexander. 2023. Estimating Local Structural Equation Models. *Journal of Intelligence* 11: 175. https://doi.org/10.3390/jintelligence11090175

Received: 4 June 2023
Revised: 16 August 2023
Accepted: 22 August 2023
Published: 1 September 2023

1. Introduction

A structural equation model (SEM) is a statistical approach for analyzing multivariate data (Bartholomew et al. 2011; Bollen 1989; Browne and Arminger 1995; Jöreskog et al. 2016; Shapiro 2012; Yuan and Bentler 2007). These models relate a multivariate vector $\boldsymbol{X} = (X_1, \ldots, X_I)$ of observed I variables (also referred to as items or indicators) to a vector of latent variables (i.e., factors) $\boldsymbol{\eta}$ of a dimension smaller than I. SEMs constrain the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of the random variable $\boldsymbol{X}$ as a function of an unknown parameter vector $\boldsymbol{\theta}$. By doing so, the mean vector is constrained as $\boldsymbol{\mu}(\boldsymbol{\theta})$, and the covariance matrix is constrained as $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.

Local structural equation models (LSEM) study SEMs as a function of a univariate moderator variable Hildebrandt et al. (2009, 2016). The moderator variable is the age or time variable in most applications. LSEM has been mentioned as a general tool for assessing measurement invariance across age or other continuous indicators in social sciences (Dong and Dumas 2020; Han et al. 2019; Leitgöb et al. 2023). Note that LSEM has also been abbreviated as LOSEM Briley et al. (2015a, 2015b).

The LSEM method is particularly suited for studying differentiation or dedifferentiation hypotheses (see Hildebrandt et al. 2009 or Molenaar et al. 2010b). Differentiation hypotheses of intelligence and general scholastic abilities describe changes in the relationship between different cognitive abilities (i.e., their structural organization) depending on the level of general ability (ability differentiation), age (differentiation in children and adolescents; dedifferentiation in older adults), and their interaction. Breit et al. (2022) presented a systematic review of 33 reports with data from 51 studies with over 260,000 participants that examined differentiation effects. The findings indicated practically significant ability differentiation in children and adults, and significant age dedifferentiation in older adults, with effect sizes that implicate a practical significance of the effects. However, Breit et al. (2022) also showed that age differentiation in children and adolescents was not supported. Instead, small but negligible effect sizes were found for age dedifferentiation in adolescents.

The LSEM method has been extended to two moderator variables by Hartung et al. (2018). Molenaar (2021) proposed a semiparametric moderated factor modeling approach in which no assumption concerning the functional form between the moderator and the model parameters are imposed. In contrast to the original definition of LSEM (Hildebrandt et al. 2009), some model parameters are allowed to be invariant across the continuous moderator variable.

LSEM is closely related to moderated nonlinear factor analysis (MNFA; Bauer 2017; Curran et al. 2014; Molenaar and Dolan 2012). In MNFA, a functional form of SEM model parameters as a function of a single moderator (or multiple moderators) is imposed. In this sense, MNFA is often more confirmatory than LSEM. Nevertheless, differentiation hypotheses were also investigated by means of MNFA (Molenaar et al. (2010a, 2010b, 2011, 2017)). A tutorial on how to apply MNFA using the R package OpenMx (Boker et al. 2011) was given by Kolbe et al. (2022). LSEM also bears a similarity to the approach of individual parameter change (Oberski 2013; Arnold et al. (2020, 2021). Variation in SEM model parameters can also be tested with score-based invariance tests (Huth et al. 2022; Merkle and Zeileis 2013; Wang et al. 2014).

LSEM has been implemented in the R package sirt (Robitzsch 2023b) as a wrapper to the popular SEM package lavaan (Rosseel 2012). Moreover, the R package umx (Bates et al. 2019) can also be utilized for LSEM estimation.

This article reviews and extends LSEM estimation methods and discusses the implementation in the R package sirt. In previous literature, LSEM was fitted as a sequence of models that are separately evaluated as each value of the moderator variables. In this article, a joint estimation approach is proposed that is a simultaneous estimation method across all moderator values and also allows some model parameters to be invariant with respect to the moderator. Sufficient detail on the core estimation functions in the sirt package is provided. The article also evaluates two significance testing approaches to assess whether the moderator values are related to a model parameter in two simulation studies. Finally, an empirical example demonstrates the usefulness of the LSEM methodology.

The remainder of this article is structured as follows. Section 2 overviews the most important LSEM applications in the literature. In Section 3, different LSEM estimation and significance testing approaches are presented. Details about LSEM implementation in the sirt package can be found in Section 4. Section 5 discusses R input code and R output of an LSEM analysis involving illustrative datasets. Section 6 includes a simulation study investigating parameter recovery in LSEM regarding bias and root mean square error. Section 7 includes a simulation study that investigates different estimators of variability in parameter curves and the statistical properties of significance tests of parameter variation. In Section 8, an empirical example is presented that reanalyzes SON-R intelligence data for children aged between $2^{1}/_{2}$ and 7 years. Finally, Section 9 closes with a discussion.

2. Review of LSEM Applications

We now review important LSEM applications to demonstrate that this method is widely applied in substantive research. The original LSEM publication of Hildebrandt et al. (2009) ("Complementary and competing factor analytic approaches for the investigation of measurement invariance") has been cited 93 times and 80 times, according to Google Scholar and ResearchGate (accessed on 18 July 2023), respectively. The second methodological LSEM publication by Hildebrandt et al. (2016) ("Exploring factor model parameters across continuous variables with local structural equation models") has been cited 111 times, 89 times, and 77 times, according to Google Scholar, ResearchGate, and Web of Science (accessed on 18 July 2023), respectively. Hence, one could say that LSEM fills some niche in the researcher's methodological toolbox.

In the following, some LSEM applications are briefly described. The studies are loosely organized according to the fields of application.

Olaru and Allemand (2022) examined differential and correlated change in personality across the adult lifespan using LSEM. Brandt et al. (2022) applied LSEM to four waves

of data obtained with the full NEO Personality Inventory collected over 11 years from 1667 adults in a US sample using age as a continuous moderator. Hartung et al. (2021) investigated the age-moderated covariance structure of the satisfaction with life scale (SWLS) and the domains of health satisfaction and financial satisfaction using LSEM. Olaru et al. (2019) analyzed NEO personality indicators across ages between 16 and 66 years by means of LSEM. They selected items for short scales that had the greatest extent of measurement invariance across age. Seifert et al. (2022) studied whether the rank-order stability of personality increases until midlife and declines later in old age and found that this inverted U-shaped pattern was not consistently observed in two reanalyzes utilizing LSEM. Loneliness across different age levels was investigated by LSEM in Entringer and Gosling (2022) and Panayiotou et al. (2022). Van den Akker et al. (2021) applied LSEM for students aged between 8 and 18 years to investigate whether levels of conscientiousness and agreeableness decrease when levels of neuroticism increase, indicating a dip in personality maturation. Gnambs (2013) applied LSEM in a multitrait multi-informant meta-analysis for the big five factors.

Hartung et al. (2022) investigated the structure of the "dark personality factor" across age and gender using LSEM. Krasko and Kaiser (2023) investigated measurement invariance across age for the dark triad by means of LSEM.

Bratt et al. (2018) investigated levels of perceived age discrimination across early to late adulthood by employing LSEM, using data from the European social survey (ESS) collected in 29 countries. Dutton and Kirkegaard (2022) applied LSEM to investigate a particular question about the association between religiousness and intelligence. Allemand et al. (2022) used LSEM to investigate the effects of continuous age and COVID-19 virus worry on mean levels and correlations between gratitude and remaining opportunities and time. Allemand et al. (2021) examined age-related psychometrics and differences in the measurement, mean-levels, variances, and correlations of gratitude and future time perspective across adulthood using data in a representative Swiss sample for participants aged between 19 and 98 years.

Schroeders et al. (2015) studied the differentiation fluid and crystallized intelligence in German students of grades 5 to 12. Watrin et al. (2022) studied the age differentiation hypothesis of declarative knowledge, as proposed in Cattell's investment theory. Hülür et al. (2011) studied with LSEM whether cognitive abilities become more differentiated with increasing age during childhood for children from age 2.5 to 7. Hartung et al. (2020) tested whether associations among executive functions strengthened from middle childhood to adolescence using cross-sectional data from a sample of children aged between 7 and 15 years. Gnambs and Schroeders (2020) examined the effects of cognitive abilities on the factor structure of the Rosenberg self-esteem scale across age by means of LSEM. Whitley et al. (2016) explored cross-sectional associations of age with five cognitive tests (word recall, verbal fluency, subtraction, number sequence, and numerical problem solving) in a large representative sample aged between 16 and 100 living in the UK. Breit et al. (2020) investigated ability differentiation, developmental differentiation, and their interaction with LSEM in two studies. Breit et al. (2021) provided a review of the literature on ability and developmental differentiation effects in children and youths. Breit et al. (2023) studied ability differentiation, including creativity measures, through LSEM for German students aged between 12 and 16 years.

Hildebrandt et al. (2010) employed LSEM to investigate structural invariance and age-related performance differences in face cognition. Hildebrandt et al. (2013) studied the specificity of face cognition compared with object cognition from individual differences and aging perspective by determining the amount of overlap between these abilities at the level of latent constructs across age. By utilizing LSEM, Liu et al. (2022) found that individual differences in white matter microstructure of the face processing brain network were more differentiated from global fibers with increasing ability.

LSEM was also applied in behavioral neurosciences Kaltwasser et al. (2017). Jokić-Begić et al. (2019) used LSEM for assessing measurement invariance across age for cyper-

chondria, a process of increased anxiety over one's health as a result of excessive online searching. Lodi-Smith et al. (2021) found that autism characteristics measured by the autism-spectrum quotient scale were not strongly associated with age by utilizing LSEM. Cox et al. (2016) used LSEM to quantify microstructural properties of the human brain's connections for understanding normal ageing and disease (see also Briley et al. (2015b)). Researchers de Mooij et al. (2018) used LSEM to study differences within and between brain and cognition across the adult life span. Zheng et al. (2019) investigated whether genetic and environmental influences on achievement goal orientations shift were moderated with age. Madole et al. (2019) applied LSEM in network analysis as a method for investigating symptom-level associations that underlie comorbidity connecting diagnostic syndromes.

Olaru et al. (2019) utilized LSEM in combination with ant colony optimization (see also Olaru and Jankowsky 2022) to resample and weight subjects to study differences in the measurement model across age as a continuous moderator variable.

An overview of different modeling strategies of LSEM for longitudinal data is presented in Olaru et al. (2020). Wagner et al. (2019) investigated through LSEM whether personality becomes more stable with age. They disentangled state and trait effects for the big five across the life span by applying LSEM to trait-state-occasion models. Gana et al. (2023) applied trait-state-occasion models in tandem with LSEM to investigate whether the characteristics of the depression EURO-D scale were associated with age.

LSEM was also applied to moderator variables different from age. Klieme and Schmidt-Borcherding (2023) employed LSEM to explore whether there is noninvariance for indicators of research self-efficacy regarding different training levels of students operationalized as the number of studied semesters. Weiss et al. (2020) investigated the threshold hypothesis of creativity by handling intelligence as a continuous moderator in LSEM. Schroeders and Jansen (2022) studied by means of LSEM whether the multidimensional structure of the science self-concept is moderated by levels of the cognitive ability in science. Basarkod et al. (2023) investigated whether reading self-concept dimensions vary across reading achievement levels in the PISA study. Olaru et al. (2022) examined the effects of family background on children's receptive vocabulary using LSEM with latent growth curve models. Bolsinova and Molenaar (2019) (see also Bolsinova and Molenaar 2018) used LSEM for indicator-specific covariates and extended LSEM to the study of cognitive tests involving reaction times.

3. Estimating and Testing Local Structural Equation Models

3.1. Single-Group Structural Equation Model

In SEM, a measurement model is imposed that relates the observed variables $\boldsymbol{X}$ to latent variables $\boldsymbol{\eta}$

$$\boldsymbol{X} = \boldsymbol{\nu} + \boldsymbol{\Lambda\eta} + \boldsymbol{\epsilon} \, . \tag{1}$$

In addition, the covariance matrix of $\boldsymbol{\epsilon}$ is denoted by $\boldsymbol{V}$; that is, $\mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi}$. Moreover, $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are multivariate normally distributed random variables. In addition, $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are assumed to be uncorrelated. In CFA, the multivariate normal (MVN) distribution is represented as $\boldsymbol{\eta} \sim \mathsf{MVN}(\boldsymbol{\alpha}, \boldsymbol{\Phi})$ and $\boldsymbol{\epsilon} \sim \mathsf{MVN}(\mathbf{0}, \boldsymbol{\Psi})$. As we are only concerned with the covariance structure in SEM in this paper, we assume $\boldsymbol{\alpha} = \mathbf{0}$ and $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\nu}$. Then, the covariance matrix of $\boldsymbol{X}$ in CFA can be computed as:

$$\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda\Phi\Lambda}^\top + \boldsymbol{\Psi} \, . \tag{2}$$

The parameter vector $\boldsymbol{\theta}$ contains parameters in $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$ that are estimated. Typically, the covariance matrix $\boldsymbol{\Sigma}$ is a constrained matrix determined by the specification (2).

In a general SEM, relationships among the latent variables $\boldsymbol{\eta}$ are modeled in path models. A matrix $\boldsymbol{B}$ of regression coefficients is specified such that:

$$\boldsymbol{\eta} = \boldsymbol{B\eta} + \boldsymbol{\zeta} \, , \tag{3}$$

where $\boldsymbol{\eta}$ denotes an endogeneous and $\boldsymbol{\zeta}$ an exogeneous multivariate normally distributed latent variables. Note that (3) can be written as:

$$\boldsymbol{\eta} = (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\zeta} \, , \tag{4}$$

where $\boldsymbol{I}$ denotes the identity matrix. In this case, the covariance matrix of $\boldsymbol{X}$ are represented in SEM as:

$$\text{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Phi}[(\boldsymbol{I} - \boldsymbol{B})^{-1}]^{\top}\boldsymbol{\Lambda}^{\top} + \boldsymbol{\Psi} \, . \tag{5}$$

Some identification constraints must be imposed when estimating the covariance structure of the SEM in (2) or (5) (Bollen 1989; Bollen and Davis 2009). The purpose of identifying constraints primarily lies in a convenient interpretation of latent variables $\boldsymbol{\eta}$ and is not primarily driven by improving the efficiency of estimating $\boldsymbol{\Sigma}$.

When modeling multivariate normally distributed data without missing data, the empirical covariance matrix $\boldsymbol{S}$ is a sufficient statistic for the unknown covariance matrix $\boldsymbol{\Sigma}$. Hence, $\boldsymbol{S}$ is also sufficient for the parameter vector $\boldsymbol{\theta}$ of the SEM in (2) or (5).

3.2. Multiple-Group Structural Equation Model

We now describe the general estimation of a multiple-group SEM. There exist G known groups $g = 1, \ldots, G$. The allocation of a group to a subject is known in this case. Assume that group g has N_g subjects and an empirical covariance matrix $\boldsymbol{S}_g$. The population covariance matrices are denoted by $\boldsymbol{\Sigma}_g$ $(g = 1, \ldots, G)$. The model-implied covariance matrices are denoted by $\boldsymbol{\Sigma}_g(\boldsymbol{\theta})$ $(g = 1, \ldots, G)$. The unknown parameter vector $\boldsymbol{\theta}$ can have common parameters across groups and parameters that are group-specific. For example, in a CFA, equal factor loadings and item intercepts across groups are frequently imposed (i.e., measurement invariance holds; Meredith 1993; Putnick and Bornstein 2016) by assuming the same loading matrix $\boldsymbol{\Lambda}$ across groups, while covariance matrices of latent variables or the matrix $\boldsymbol{B}$ of regression coefficients are allowed to differ across groups.

Up to constants, the maximum likelihood (ML) fitting function of the unknown parameter $\boldsymbol{\theta}$ for the covariance structure in the multiple-group SEM is given by (see Bollen 1989 and Jöreskog et al. 2016):

$$F(\boldsymbol{\theta}; \{\boldsymbol{S}_g\}_g) = \sum_{g=1}^{G} N_g \Big(\log|\boldsymbol{\Sigma}_g(\boldsymbol{\theta})| + \text{tr}(\boldsymbol{S}_g\boldsymbol{\Sigma}_g(\boldsymbol{\theta})^{-1}) - \log|\boldsymbol{S}_g| - I \Big) \, . \tag{6}$$

Note that I refers to the number of observed variables; that is, the dimension of $\boldsymbol{X}$. The set $\{\boldsymbol{S}_g\}_g$ denotes the set of G empirical covariance matrices that are sufficient statistics in multiple-group SEM estimation. The parameter vector $\boldsymbol{\theta}$ is estimated by minimizing F in (6) and is denoted as the ML estimate. The estimated parameter is denoted by $\hat{\boldsymbol{\theta}}$.

In practice, the model-implied covariance matrix can be misspecified (Boos and Stefanski 2013; Gourieroux et al. 1984; Kolenikov 2011; White 1982), and $\boldsymbol{\theta}$ is a pseudo-true parameter defined as the minimizer of the fitting function F in (6). Importantly, $\boldsymbol{\theta}$ does not refer to a parameter of the data-generating model in this case. In contrast, it should be interpreted as a summary of the data that are of central interest to the researcher.

The ML fitting function (6) can be considered a special case of discrepancy function. To this end, we define a general discrepancy function $\mathcal{D}(\boldsymbol{S}, \boldsymbol{\Sigma})$ between an empirical covariance matrix $\boldsymbol{S}$ and a population covariance matrix $\boldsymbol{\Sigma}$. The real-valued nonnegative function $\mathcal{D}$ should only attain the value zero if $\boldsymbol{S} = \boldsymbol{\Sigma}$ (i.e., for correctly specified models). For the ML fitting function, the discrepancy function $\mathcal{D}$ is defined as:

$$\mathcal{D}(\boldsymbol{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| + \text{tr}(\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}) - \log|\boldsymbol{S}| - I \, . \tag{7}$$

Using definition (7), we can rewrite (6) as:

$$F(\boldsymbol{\theta}; \{\boldsymbol{S}_g\}_g) = \sum_{g=1}^{G} N_g \mathcal{D}(\boldsymbol{S}_g, \boldsymbol{\Sigma}_g(\boldsymbol{\theta})) \ , \tag{8}$$

and $\hat{\boldsymbol{\theta}}$ is the minimizer of $F(\boldsymbol{\theta}; \{\boldsymbol{S}_g\}_g)$.

If an age moderator variable A is available, an SEM can, in principle, be estimated for all subgroups of subjects for different values of the age variable. In practice, sample sizes for concrete age values might be too small for separate estimation of the SEM. Moreover, discretizing the values of a continuous moderator variable A into G distinct groups of subjects might not be preferred due to loss of information (Hildebrandt et al. 2009). To circumvent these issues, LSEM has been proposed. We discuss LSEM estimation methods in the next subsections.

3.3. Local Weighting

Instead of grouping subjects that fall within a given range of the moderator, as in multiple-group SEMs, observations are locally weighted around focal points (i.e., specific values of the continuous moderator variable) in LSEM. In previous studies, SEMs are sequentially estimated on the basis of weighted samples of observations at all focal points (i.e., the pointwise LSEM estimation approach, see Section 3.5).

In LSEM, researchers are interested in investigating moderator-specific covariance structures. That is, they aim to model conditional covariances:

$$\text{Var}(\boldsymbol{X}|A = a) = \boldsymbol{\Sigma}(a) \tag{9}$$

As argued in the previous section, sample sizes might be too small for estimating $\boldsymbol{\Sigma}(a)$ only for subjects with $A = a$. To this end, subjects with moderator values a sufficiently close to a focal point a_t (i.e., a chosen value of the moderator variable A) should also enter the estimation. For each focal point a_t and each subject n, weights w_{nt} are computed that reflect the distance of the moderator value (e.g., a value of age) of person n (i.e., a_n) and the focal point a_t. If $a_n = a_t$, the weight should be one, and it should be zero for age values a_n that strongly differ from a_t.

The computation of weights relies on a kernel function K that is chosen by the researcher (Hildebrandt et al. (2009, 2016). The real-valued kernel function fulfills the properties $K(0) = 1$, $K(x) = K(-x)$ (i.e., it is a symmetry function), $K(x) \geq 0$ for all $x \in \mathbb{R}$, and K is a decreasing function for $x \geq 0$. The subject-specific weight w_{nt} for subject n at a focal point a_t with a pre-specified bandwidth bw is computed as:

$$w_{nt} = K\left(\frac{a_n - a_t}{bw}\right) . \tag{10}$$

By the definition of K, weights are bounded within the interval $[0, 1]$.

Typical choices of the weight function in the literature of nonparametric regression or density estimation are the Gaussian kernel, the Epanechnikov kernel, and the uniform kernel function. The Gaussian kernel function is defined as:

$$K(x) = \exp(-x^2/2) \ . \tag{11}$$

In density estimation involving the Gaussian kernel function, an optimal bandwidth is given by $bw = hN^{-1/5}\sigma_A$ with $h = 1.1$, and σ_A is the standard deviation of the age moderator variable (Silverman 1986). The parameter h is referred to as the bandwidth factor in this article. The Epanechnikov kernel function is defined as:

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2) & \text{for } |x| \leq 1 \\ 0 & \text{for } |x| > 1 \end{cases} \ . \tag{12}$$

For age values a_n with $|a_n - a_t| > bw$, weights w_{nt} are zero. Finally, the uniform kernel function is defined as:

$$K(x) = \begin{cases} 1 & \text{for } |x| \leq 1 \\ 0 & \text{for } |x| > 1 \end{cases} . \tag{13}$$

The uniform kernel can be used to define weights so that they reflect the discretization of the continuous age variable A into G distinct groups. The estimated LSEM will provide parameter results that are identical to the multiple-group SEM if the same identification constraints are utilized.

3.4. Estimation of Conditional Means and Conditional Covariances

We now describe the estimation conditional covariances $\boldsymbol{\Sigma}(a)$. In a practical implementation of the LSEM, researchers define a discrete grid of moderator values $a_1, a_2, \ldots, a_T$ (i.e., the focal points) of the age variable A. In most applications, a grid of equidistant focal points is chosen (Hildebrandt et al. 2016). However, the grid of focal points could also be chosen in such a way that it mimics the empirical distribution of the moderator variable. For example, researchers might use empirical percentiles of the moderator variable (e.g., a grid of 10 focal points using the pth percentile for $p = 5, 15, \ldots, 95$).

To estimate conditional covariances at a focal point a_t, we first compute the conditional mean function $\mathrm{E}(\boldsymbol{X}|A = a)$ for $\boldsymbol{X} = (X_1, \ldots, X_I)$. For a variable X_i for $i = 1, \ldots, I$, a local quadratic regression model is specified to estimate the conditional mean at focal point a_t. That is, one minimizes:

$$(\hat{\gamma}_{it0}, \hat{\gamma}_{it1}, \hat{\gamma}_{it2}) = \underset{(\gamma_{it0}, \gamma_{it1}, \gamma_{it2})}{\arg\min} \left\{ \sum_{n=1}^{N} w_{nt} \Big(x_{in} - \gamma_{it0} - \gamma_{it1}(a_{nt} - a_t) - \gamma_{it2}(a_{nt} - a_t)^2 \Big)^2 \right\} \tag{14}$$

The conditional mean estimate of $\mu_i(a_t) = \mathrm{E}(X_i|A = a_t)$is given by $\hat{\mu}_i(a_t) = \hat{\gamma}_{it0}$. Note that the minimization in (14) is a weighted least squares estimation problem for a linear regression (i.e., it is linear in model parameters) and closed formulae are available for estimating $(\gamma_{it0}, \gamma_{it1}, \gamma_{it2})$ (see Fox 2016).

We now describe the estimation of conditional covariances $\sigma_{ij}(a_t) = \mathrm{Cov}(X_i, X_j|A = a_t)$. First, residuals e_{nit} are computed using local quadratic regression parameters defined in (14) as:

$$e_{nit} = x_{in} - \hat{\gamma}_{it0} - \hat{\gamma}_{it1}(a_{nt} - a_t) - \hat{\gamma}_{it2}(a_{nt} - a_t)^2 . \tag{15}$$

The estimate of the conditional covariances $\sigma_{ij}(a_t)$ can be obtained by simple weighting or a local regression model.

In the weighting approach, one estimates:

$$\hat{\sigma}_{ij}(a_t) = W_t^{-1} \sum_{n=1}^{N} w_{nt} e_{nit} e_{njt} , \tag{16}$$

where $W_t = \sum_{n=1}^{N} w_{nt}$. This approach was advocated in Hildebrandt et al. (2009) and Hildebrandt et al. (2016).

In recently proposed local regression modeling (see Olaru et al. 2020), one also specifies a local quadratic regression estimation problem for the computation of the conditional covariance:

$$(\hat{\delta}_{ijt0}, \hat{\delta}_{ijt1}, \hat{\delta}_{ijt2}) = \underset{(\delta_{ijt0}, \delta_{ijt1}, \delta_{ijt2})}{\arg\min} \left\{ \sum_{n=1}^{N} w_{nt} \Big(e_{nit} e_{njt} - \delta_{ijt0} - \delta_{ijt1}(a_{nt} - a_t) - \delta_{ijt2}(a_{nt} - a_t)^2 \Big)^2 \right\} . \tag{17}$$

The estimate of the conditional covariance is given as $\hat{\sigma}_{ij}(a_t) = \hat{\delta}_{ijt0}$.

Note that the estimation of the conditional mean function in (14) and the conditional covariance function in (17) is essentially equivalent, except for the case that the former uses the values x_{ni} as the dependent variable x_{ni} (i.e., indicator i), while the latter uses the product residual $e_{nit}e_{njt}$ of variables for indicators i and j for the computation of the moderator-specific conditional covariance.

The steps can be repeated for all pairs of variables i and j ($i, j = 1, \ldots, I$) and all focal points a_t ($t = 1, \ldots, T$). The resulting estimated conditional covariance matrices at focal points a_t are denoted by $\hat{\boldsymbol{\Sigma}}_t$ ($t = 1, \ldots, T$). The estimated covariance matrices $\hat{\boldsymbol{\Sigma}}_t$ are not guaranteed to be positive definite. Therefore, the estimate might be slightly modified to determine a close matrix to $\hat{\boldsymbol{\Sigma}}_t$ that fulfills the positive definiteness property (Bentler and Yuan 2011).

LSEM estimation methods rely on the estimated conditional covariances. Three different estimation approaches are described in Sections 3.5, 3.6 and 3.8.

3.5. Pointwise LSEM Estimation

Pointwise LSEM estimation relies on the idea that a separate SEM is fitted to each focal point a_t. The resulting parameter estimates $\hat{\boldsymbol{\theta}}_t$ are plotted or analyzed as a function of the age variable A. More formally, based on the conditional covariance estimate $\hat{\boldsymbol{\Sigma}}_t$, at each focal point a_t, the following fitting function is minimized:

$$F(\boldsymbol{\theta}_t; \hat{\boldsymbol{\Sigma}}_t) = \mathcal{D}\big(\hat{\boldsymbol{\Sigma}}_t, \boldsymbol{\Sigma}_t(\boldsymbol{\theta}_t)\big) \ , \tag{18}$$

where $\hat{\boldsymbol{\theta}}_t$ denotes the minimizer of $F(\boldsymbol{\theta}_t; \hat{\boldsymbol{\Sigma}}_t)$. Note that in (18), the distance between the empirical conditional covariance $\hat{\boldsymbol{\Sigma}}_t$ and the model-implied conditional covariance $\boldsymbol{\Sigma}_t(\boldsymbol{\theta}_t)$ at the focal point a_t is minimized. This approach was proposed by Hildebrandt et al. (2009, 2016). The minimization in (18) is not restricted to ML estimation and can also be applied to weighted least estimation in SEM (Browne 1974) or model-robust fitting functions (Robitzsch 2023a).

Model fit statistics, such as RMSEA, SRMR, or TLI, are computed at each value of the focal point. Note that pointwise LSEM estimation provides parameter curves across different values of the moderator variable.

The pointwise LSEM estimation method allows the parameter vector $\boldsymbol{\theta}(a)$ to vary freely across a. However, this flexibility sometimes hinders interpretation. Moreover, some researchers might prefer to impose invariance constraints for some of the model parameters (Leitgöb et al. 2023). For this reason, a joint LSEM estimation approach is proposed that is described in the next Section 3.6.

3.6. Joint LSEM Estimation with Invariance Constraints

While pointwise LSEM estimation tackles the estimation problem by successively and separately estimating an SEM at each of the focal points, joint LSEM estimation defines a single estimation function that involves conditional covariance matrices of all focal points. By doing so, the parameter vector $\boldsymbol{\theta}$ can contain parameters that are specific to each focal point and parameters that do not vary for different values of age. The fitting function is defined as:

$$F(\boldsymbol{\theta}; \{\hat{\boldsymbol{\Sigma}}_t\}_t) = \sum_{t=1}^{T} W_t \mathcal{D}(\hat{\boldsymbol{\Sigma}}_t, \boldsymbol{\Sigma}_t(\boldsymbol{\theta})) \ , \tag{19}$$

where $\hat{\boldsymbol{\theta}}$ is the minimizer of $F(\boldsymbol{\theta}; \{\hat{\boldsymbol{\Sigma}}_t\}_t)$ and $W_t = \sum_{n=1}^{N} w_{nt}$ is the sum of weights specific to each focal point a_t. Note that (19) looks like a fitting function in multiple-group SEM estimation. However, subjects can enter multiple groups (i.e., focal points) because they enter the estimated conditional covariances multiple times according to the weights w_{nt}. Hence, the fitting function F in (19) will not be an ML fitting function and falls in the general class of M-estimation problems (Stefanski and Boos 2002).

The parameter vector $\boldsymbol{\theta}$ can be decomposed into components $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T)$, where $\boldsymbol{\theta}_0$ contains parameters that are invariant across age, and $\boldsymbol{\theta}_t$ for $t \geq 1$ contain the parameters that vary across age values. The fitting function in (19) can then be rewritten as:

$$F(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T; \{\hat{\boldsymbol{\Sigma}}_t\}_t) = \sum_{t=1}^{T} W_t \mathcal{D}(\hat{\boldsymbol{\Sigma}}_t, \boldsymbol{\Sigma}_t(\boldsymbol{\theta}_0, \boldsymbol{\theta}_t)) \ . \tag{20}$$

Note that the originally proposed pointwise estimation of the fitting function in (18) is equivalent to joint LSEM estimation in (20) if there does not exist invariant model parameters $\boldsymbol{\theta}_0$.

In joint LSEM estimation, global model fit statistics are computed. These fit statistics can be interpreted similarly as in multiple-group SEMs.

3.7. Estimation of DIF Effects

In joint LSEM estimation defined by the fitting function F in (20), some parameters (i.e., the parameter vector $\boldsymbol{\theta}_0$) have invariance constraints across the age moderator variable. These invariance constraints ease interpretation and have the advantage of specifying parsimonious SEMs. However, researchers might be interested in what would happen if these invariance constraints were freed.

Violations of measurement invariance are referred to as differential item functioning (DIF) in item response theory literature (Mellenbergh 1989; Holland and Wainer 1993; Millsap 2011). Noninvariant parameters are referred to as DIF effects in this literature. We also use this notation and now discuss the estimation of DIF effects. DIF effects emerge if all estimated age-specific parameters $\hat{\boldsymbol{\theta}}_t$ ($t \geq 1$) are held fixed in (20), and the entries of the parameter vector $\boldsymbol{\theta}_0$ are allowed to vary across age. We denote the focal-point-specific estimates of DIF effects by $\boldsymbol{\delta}_t$. To this end, invariant parameters $\boldsymbol{\theta}_0$ are replaced with $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_T$, and the following fitting function F is minimized to obtain DIF effect estimates $\hat{\boldsymbol{\delta}}_t$ ($t = 1, \ldots, T$):

$$F(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_T, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_T; \{\hat{\boldsymbol{\Sigma}}_t\}_t) = \sum_{t=1}^{T} W_t \mathcal{D}(\hat{\boldsymbol{\Sigma}}_t, \boldsymbol{\Sigma}_t(\boldsymbol{\delta}_t, \hat{\boldsymbol{\theta}}_t)) \; . \tag{21}$$

Note that there are no invariant model parameters in (21), and the DIF effects $\boldsymbol{\delta}_t$ at the focal point a_t could alternatively be obtained by pointwise minimization of:

$$F(\boldsymbol{\delta}_t, \hat{\boldsymbol{\theta}}_t; \hat{\boldsymbol{\Sigma}}_t) = \mathcal{D}(\hat{\boldsymbol{\Sigma}}_t, \boldsymbol{\Sigma}_t(\boldsymbol{\delta}_t, \hat{\boldsymbol{\theta}}_t)) \; . \tag{22}$$

The estimated DIF effects can be plotted or analyzed as a function of the age moderator to investigate whether the invariance constraints are substantially violated.

3.8. Joint LSEM Estimation with More General Parameter Constraints and Relation to Moderated Nonlinear Factor Analysis

In this subsection, joint LSEM estimation is slightly generalized. The fitting function is the same as in (20), but constrains across focal-point-specific parameters $\boldsymbol{\theta}_t$ are allowed. In particular, we discuss the implementation of linear, quadratic, and piecewise linear or quadratic parameter constraints.

Assume a parameter curve $\theta(a_t)$ for a particular parameter. Furthermore, assume that the focal points are equidistant; that is, $a_{t+1} - a_t = \Delta$ are equal for $t = 1, \ldots, T-1$.

We first describe a linear parameter constraint. A linear function of a parameter θ for age values a is given by $f(a) = \alpha_0 + \alpha_1 a$. The first derivative of f is constant, and it holds that $f'(a_{t+1}) = f'(a_t) = \alpha_1$. Hence, the equality in derivatives can be translated into equalities in first-order differences in model parameters:

$$\theta(a_{t+2}) - \theta(a_{t+1}) = \theta(a_{t+1}) - \theta(a_t) \; . \tag{23}$$

These constraints can be added in multiple-group SEM in typical SEM software such as lavaan (Rosseel 2012).

A quadratic function of a parameter is given by $f(a) = \alpha_0 + \alpha_1 a + \alpha_2 a^2$. This function has constant second-order derivatives; that is, $f''(a_{t+1}) = f''(a_t) = 2\alpha_2$. Hence, second-order differences in parameter values are constant, which translates into:

$$\theta(a_{t+3}) - 2\theta(a_{t+2}) + \theta(a_{t+1}) = \theta(a_{t+2}) - 2\theta(a_{t+1}) + \theta(a_t) \; . \tag{24}$$

Similarly, cubic parameter constrains can be implemented by recognizing that the third-order differences in parameter values are constant. A slightly more tedious constraint than (24) can be derived.

The linearity and quadratic constraints in (23) and (24) can also be applied if parameter curves are broken into segments. Hence, piecewise linear or quadratic functions can be applied.

Applying (piecewise) quadratic parameter functions in joint LSEM estimation can be interpreted as a kind of smoothing procedure to stabilize parameter estimation. Furthermore, the raw data are smoothed when computing the estimated conditional covariance matrices $\hat{\boldsymbol{\Sigma}}_t$. Hence, researchers have two choices for how stabilizing parameter estimation in LSEM.

Notably, parameter constraints in joint LSEM estimation are estimates of MNFA in a particular case. If the age moderator values A has only values at the grid of equidistant focal points $a_1, \ldots, a_T$, then using the uniform kernel with $bw = (a_2 - a_1)/2$ is equivalent to MNFA with appropriate parameter constraints. Such an approach is described in Tucker-Drob (2009).

3.9. Parameter Curve Summaries and Significance Testing

Finally, we discuss the definition of summary statistics and the test of significant parameter variation across age. Let $\theta(a_t)$ be a parameter curve of some model parameter estimated at focal points a_t ($t = 1, \ldots, T$). The parameter curve $\theta(a)$ can be summarized by the mean and the standard deviation. Let $f(a_t)$ be the discrete density of the age variable A at focal point a_t and assume that $\sum_{t=1}^{T} f(a_t) = 1$. The (weighted) average value of the parameter curve (i.e., the mean) is given as:

$$\mathrm{M}_{\theta(a)} = \sum_{t=1}^{T} f(a_t)\theta(a_t) \; . \tag{25}$$

In practice, an estimate of (25) is obtained by

$$\widehat{\mathrm{M}}_{\theta(a)} = \sum_{t=1}^{T} \hat{f}(a_t)\hat{\theta}(a_t) \; . \tag{26}$$

The standard deviation of a parameter curve quantifies the variability of a parameter curve across age and is given by:

$$\mathrm{SD}_{\theta(a)} = \sqrt{\sum_{t=1}^{T} f(a_t)\Big(\theta(a_t) - \mathrm{M}_{\theta(a)}\Big)^2} \; . \tag{27}$$

An estimate of the standard deviation defined in (27) is given by:

$$\widehat{\mathrm{SD}}_{\theta(a)} = \sqrt{\sum_{t=1}^{T} \hat{f}(a_t)\Big(\hat{\theta}(a_t) - \widehat{\mathrm{M}}_{\theta(a)}\Big)^2} \; . \tag{28}$$

The sample estimate $\widehat{\mathrm{SD}}_{\theta(a)}$ is always positive in finite samples if no invariance constraints are imposed. Hence, the naive standard deviation estimate in (28) will be positively biased. The bootstrap resampling procedure (Efron and Tibshirani 1994) can be used to reduce the bias in an estimate of $\mathrm{SD}_{\theta(a)}$. For LSEM, nonparametric bootstrap is implemented, which resamples subjects with replacement. The pointwise standard deviation of a parameter value across bootstrap samples can be used as a standard error estimate. A bias-corrected estimate of the standard deviation is obtained by:

$$\widehat{\mathrm{SD}}_{\theta(a),\mathrm{bc}} = \mathrm{sqrt}_{+}\left(\widehat{\mathrm{SD}}^2_{\theta(a)} - \mathcal{B}_{\theta(a)}\right) \; , \tag{29}$$

where $\text{sqrt}_+(x) = \sqrt{\max(x,0)}$ and $\mathcal{B}_{\theta(a)}$ is the finite-sample bias of $\widehat{\text{SD}}^2_{\theta(a)}$ that can be determined by bootstrap resampling (Efron and Tibshirani 1994). A t-statistic for significant variation in an estimated parameter curve can be computed as:

$$t = \widehat{\text{SD}}_{\theta(a),\text{bc}} / SE \,, \tag{30}$$

where SE is the standard deviation of $\widehat{\text{SD}}_{\theta(a)}$ values defined in (28) across different bootstrap samples. Note that this test procedure relies on a normal distribution assumption for the test statistic t, although it is probably an incorrect null distribution.

An alternative test for parameter variation is based on a Wald test. A covariance matrix estimate $\boldsymbol{V}$ for the vector $\boldsymbol{\xi} = (\theta(a_1), \ldots, \theta(a_T))$ can be obtained from bootstrap. It is assumed that $\hat{\boldsymbol{\xi}}$ is multivariate normally distributed. Let $\boldsymbol{H}$ be a $(T-1) \times T$ matrix that implements equality constraints across the values of the parameter curve. The null hypothesis of no parameter variation is given by $\boldsymbol{H\xi} = \boldsymbol{0}$. Consider the Wald test statistic:

$$\chi^2 = \hat{\boldsymbol{\xi}}^\top \boldsymbol{H}^\top \left(\boldsymbol{H}^\top \boldsymbol{V} \boldsymbol{H} \right)^{-1} \boldsymbol{H} \hat{\boldsymbol{\xi}} \tag{31}$$

This statistic is chi-square distributed with $T-1$ degrees of freedom.

In previous work, a permutation test has been proposed for testing parameter variation (Hartung et al. 2022; Hildebrandt et al. (2009, 2016)). A permutation test simultaneously assesses the effects on all parameters. In contrast, the test based on the standard deviation (30) and the Wald test (31) relies on a fitted model without modifying all other model parameters. Hence, we tend to favor the latter statistics over the permutation test.

4. Implementation of Local Structural Equation Models in the Sirt Package

In this section, we discuss the implementation of LSEM in the R (R Core Team 2023) package sirt (Robitzsch 2023b). The CRAN version can be installed within R using `utils::install.packages('sirt')`, while the most recent GitHub version can be installed employing `devtools::install_github('alexanderrobitzsch/sirt')`. The four primary LSEM functions are `sirt::lsem.estimate()`, `sirt::lsem.bootstrap()`, `sirt::lsem.test()` and `sirt::lsem.permutationTest()`, which will be discussed below. The new CRAN release of sirt from August 2023 (sirt 3.13-228; https://cran.r-project.org/web/packages/sirt/ accessed on 11 August 2023) includes the functionality described in this article.

LSEM estimation in sirt provides a wrapper to the SEM package lavaan (Rosseel 2012). The model specification follows the lavaan syntax, which eases the familiarity with R code for LSEM estimation because lavaan seems to be the most popular open-source SEM software.

In Listing 1, the main function `sirt::lsem.estimate()` is displayed. This function is the main LSEM estimation function. We now discuss the most important arguments in detail.

Listing 1. LSEM function `sirt::lsem.estimate()`.

```
sirt::lsem.estimate(data, moderator, moderator.grid, lavmodel, type="LSEM", h=1.1, bw=NULL,
residualize=TRUE, fit_measures=c("rmsea","cfi","tli","gfi","srmr"),
standardized=FALSE, standardized_type="std.all", lavaan_fct="sem",
sufficient_statistics=TRUE, pseudo_weights=0, sampling_weights=NULL,
loc_linear_smooth=TRUE, est_joint=FALSE,
par_invariant=NULL, par_linear=NULL, par_quadratic=NULL,
partable_joint=NULL, pw_linear=1, pw_quadratic=1, pd=TRUE, est_DIF=FALSE,
se=NULL, kernel="gaussian", eps=1E-8, verbose=TRUE, ... )
```

In `data`, a data frame must be provided by the user. The data frame should also include the moderator variable, whose variable name must be specified in `moderator`. The set of focal points can be defined as a vector `moderator.grid`. In `lavmodel`, lavaan syntax must be

provided for estimating the LSEM. The default of the argument `type` is ‘‘`LSEM`’’; that is, an LSEM is estimated. By choosing `type=''MGM''`, a multiple-group model with a discretized moderator variable is estimated. The bandwidth in `sirt::lsem.estimate()` can be specified by `h` or `bw`. The arguments are related through the formula:

$$bw = hN^{-1/5}\hat{\sigma}_A \, , \tag{32}$$

where $\hat{\sigma}_A$ denotes the estimated standard deviation of the moderator variable A (i.e., the argument `moderator`). The logical argument `residualize` indicates whether local regression smoothing of the mean structure should be applied before estimating conditional covariances. The argument `fit_measures` defines fit statistics available in lavaan that should be included in the LSEM output. The logical argument `standardized` defines whether standardized parameters should appear in the LSEM output. The type of standardization is specified in `standardized_type` whose conventions follow the lavaan package. In `lavaan_fct`, the lavaan function is specified that is used for LSEM estimation. The default `lavaan_fct="sem"` refers to `lavaan::sem()`. Other options are `"cfa"` (for `lavaan::cfa()`) and `"lavaan"` (for `lavaan::lavaan()`). The logical argument `sufficient_statistics` indicates whether sufficient statistics (i.e., conditional mean and conditional covariances) should be used in estimation. Without missing data, ML can always rely on sufficient statistics. However, in the presence of missing data, conditional covariance matrices are estimated based on pairwise deletion. However, if full information maximum likelihood was utilized, the mean structure cannot be properly residualized. Hence, researchers are advised either to believe in missing data mechanisms close to missing completely at random that justify the usage of pairwise deletion or to apply an appropriate multiple imputation procedure prior to LSEM analysis if there are missing values in the dataset.

Users can also input a vector of sampling weights in `sampling_weights`. The logical argument `loc_linear_smooth` defines whether local quadratic regression (see (17)) should be applied in the estimation of conditional covariances. If the default `loc_linear_smooth=TRUE` is changed into `loc_linear_smooth=FALSE`, the weighting formula (16) is utilized. The logical argument `est_joint` indicates whether joint LSEM estimation (i.e., the default; see Sections 3.6 or 3.8) or pairwise LSEM estimation (see Section 3.5) is applied. Invariant model parameters can be specified in the vector argument `par_invariant`. If there are some invariant parameters, joint LSEM estimation is automatically chosen (i.e., `est_joint=TRUE`). Linear or quadratic parameter constraints on model parameters (see Section 3.8) can be specified with `par_linear` and `par_quadratic`, respectively. The number of segments in piecewise linear or piecewise quadratic parameter constrained estimation can be specified with `pw_linear` or `pw_quadratic`. The default is that the constrains should be applied across all moderator values (i.e., there is only one segment of a piecewise linear or quadratic function). The argument `partable_joint` allows the input of a lavaan parameter table in joint estimation. This argument has the advantage that arbitrary parameter constraints can be specified by the user (e.g., additional equality constraints in piecewise quadratic functions). The logical argument `pd` indicates whether non-positive definite conditional covariance matrices should be smoothed to ensure positive definiteness. The logical argument `est_DIF` defines whether DIF effects should be estimated (see Section 3.7). Note that DIF effects can only be estimated if the LSEM model contains some invariant model parameters. The argument `kernel` allows the choice of the kernel function. Possible options are ‘‘`gaussian`’’, ‘‘`epanechnikov`’’, and ‘‘`uniform`’’. Finally, the logical argument `verbose` indicates whether some output should be displayed in the R console when estimating the LSEM model.

Listing 2 displays the LSEM bootstrapping function in the sirt package. An object `object` must be provided that is the output of the `sirt::lsem.estimate()` function. The number of bootstrap samples can be specified by the argument `R`. Bootstrap can also be applied at the level of higher-order units. For example, school classes, schools, or organizations can be bootstrapped instead of bootstrapping subjects. Such a kind of cluster bootstrap is required if there is an additional dependency structure in the data. In this case, users can define a vector of cluster units in `cluster`. The `sirt::lsem.bootstrap()` also

allows more general replication designs such as jackknife, balanced repeated replication, or half sampling (Kolenikov 2010) by providing an $N \times R$ matrix of resampling weights in the argument `repl_design`. In the case of more complex designs, a scale factor `repl_factor` must be defined by the user for a correct standard error computation. In the case of jackknife, it is 1 (or $(R-1)/R$), while it is $1/R$ in the case of bootstrap resampling. The bootstrap function `sirt::lsem.bootstrap()` is needed for computing the standard deviation statistic of parameter curves and its statistical inference (see Section 3.9). The `sirt::lsem.bootstrap()` function also allows an option for parallel computing. The number of employed cores can be specified by the argument `n.core`. The default is the use of one core which means that no parallel computing is applied in LSEM bootstrap estimation.

Listing 2. LSEM function `sirt::lsem.bootstrap()`.

```
sirt::lsem.bootstrap(object, R=100, verbose=TRUE, cluster=NULL,
repl_design=NULL, repl_factor=NULL, use_starting_values=TRUE,
n.core=1, cl.type="PSOCK")
```

Listing 3 displays the LSEM function `sirt::lsem.test()` that performs the Wald tests for parameter variation (see Section 3.9). Instead of applying a test of the equality of a parameter curve on T focal points $a_1, \ldots, a_T$, the specification in `models` allows the test of significant regression parameters for a particular function. For example, a specification `"FX=~X1"=y~m+I(m^2)` tests whether the vector of the linear and the quadratic regression coefficient of the factor loading `FX=~X1` differs from $(0, 0)$. Note that `sirt::lsem.test()` requires the output of `sirt::lsem.estimate()` in `mod` and the output of the application of the bootstrap (or general resampling) of `sirt::lsem.bootstrap()` in `bmod`.

Listing 3. LSEM function `sirt::lsem.test()`.

```
sirt::lsem.test(mod, bmod, models=NULL )
```

Listing 4 displays the LSEM function `sirt::lsem.permutationTest()` that carries out the permutation test for a statistical significance test for variation in parameter curves of the LSEM model Hildebrandt et al. (2009, 2016). In the permutation test, the values of the moderator variables are randomly resampled in the dataset to create a null distribution of parameter curves under the assumption of no relation to the moderator. The number of permutation samples can be specified in the argument `B`. As in `sirt::lsem.bootstrap()`, parallel computing can be requested by the number of cores in the argument `n.core`.

Listing 4. LSEM function `sirt::lsem.permutationTest()`.

```
sirt::lsem.permutationTest(lsem.object, B=1000, residualize=TRUE, verbose=TRUE,
n.core=1, cl.type="PSOCK")
```

5. Illustrative Datasets

In this section, we illustrate LSEM estimation with the R package sirt. Three simulated datasets involving six variables `X1`, `X2`, `X3`, `Y1`, `Y2`, and `Y3` are used for illustration. The analysis model is a two-dimensional factor model with a simple loading structure, where the first factor `FX` is measured by `X1`, `X2`, and `X3`, and the second factor `FY` is measured by `Y1`, `Y2`, and `Y3`. The moderator variable age was assessed at 13 time points, referring to ages $6, 7, \ldots, 18$. An anonymous reviewer pointed out that using 13 time points would look like longitudinal data. However, we only used the 13 time points for illustratory purposes. For example, there could be 13 cross-sectional age groups that are assessed.

The population parameters of the factor model for each age $a = 6, 7, \ldots, 18$ and each of the three datasets *DATA1*, *DATA2*, and *DATA3* can be found in the directory

"*POPPARS*" at https://osf.io/puaz9/?view_only=63ffb2fd30f5400e89c59d03366bf793 (accessed on 3 June 2023). From these population parameters, 10,000 subjects were simulated at each of the 13 age points. The distribution at each age point exactly coincides with the specified conditional mean vector and the conditional covariance matrix (see, e.g., the `lavaan::simulateData()` function with the argument `empirical=FALSE` for a similar functionality). Data were simulated from a multivariate normal distribution. This simulation ensures that the population data involving 130,000 subjects (i.e., $= 13 \times 10,000$ subjects) exactly follows the specified covariance structure. In *DATA1*, all model parameters except for residual variances were assumed noninvariant. In *DATA2*, only the structural parameters (i.e., factor correlation and factor variances) were noninvariant, while factor loadings and residual variances were assumed invariant. In *DATA3*, all measurement and structural model parameters were assumed invariant. The population datasets and the data-generating model parameters can be found in the directory "*POPDATA*" at https://osf.io/puaz9/?view_only=63ffb2fd30f5400e89c59d03366bf793 (accessed on 3 June 2023). The illustrative datasets used in this section were subsamples of 2000 subjects from datasets *DATA1*, *DATA2*, and *DATA3*. The main motivation for using a subsample of the data is to show that LSEM produces some variability in model parameter estimates even if the model parameter is invariant across the moderator values in the data-generating model. The subsamples were created by random sampling without replacement from the population datasets. These datasets can be found in the directory "*ILLUSDATA*" at https://osf.io/puaz9/?view_only=63ffb2fd30f5400e89c59d03366bf793 (accessed on 3 June 2023).

Listing 5 contains the specification of the LSEM model involving two factors `FX` and `FY`. In lines 5–10 in Listing 5, the lavaan syntax for the factor model is specified in the string `lavmodel`. Line 13 in Listing 5 defines the parameter names (i.e., the factor loadings of `X2`, `X3`, `Y2`, and `Y3`) that are assumed invariant across the values of the moderator variable age. Line 16 in Listing 5 specifies the vector of focal points at which the LSEM model should be estimated. Lines 19–21 in Listing 5 contain the R command for applying `sirt::lsem.estimate()`. Note that the invariant model parameters are provided with the argument `par_invariant`, DIF effects were estimated due to `est_DIF=TRUE`, and the bandwidth factor h was chosen as 1.1. Joint LSEM estimation was applied because invariance constraints among parameters were imposed. In line 25 in Listing 5, the random seed is fixed, which ensures that bootstrap resampling will not change when applying code at a different time. Line 26 in Listing 5 specifies bootstrapping using `sirt::lsem.bootstrap()`. In total, $R = 200$ bootstrap samples were utilized. Note that the specified factor model in Listing 5 is misspecified for the dataset *DATA1*, but correctly specified for the datasets *DATA2* and *DATA3*.

Listing 5. Illustrative datasets: Specification of LSEM with invariant factor loadings in `sirt::lsem.estimate()` and subsequent bootstrap in `sirt::lsem.bootstrap()`.

```
library(lavaan)
library(sirt)

# specify model using lavaan syntax
lavmodel <- "
FX=~ 1*X1+X2+X3
FY=~ 1*Y1+Y2+Y3
FX ~~ FX
FY ~~ FY
FX ~~ FY"

#- define invariant parameters
par_invariant <- c("FX=~X2","FX=~X3","FY=~Y2","FY=~Y3" )

#- define grid of moderator values
moderator.grid <- seq(6,18,1)

# estimate LSEM model
mod <- sirt::lsem.estimate(dat, moderator="age", moderator.grid=moderator.grid,
sufficient_statistics=TRUE, lavmodel=lavmodel, h=1.1,
par_invariant=par_invariant, standardized=TRUE, est_DIF=TRUE)
summary(mod)  # print summary

# perform bootstrap with R=200 bootstrap samples
set.seed(789)
rmod <- sirt::lsem.bootstrap(mod, R=200)
summary(rmod)
```

A part of R output of the `sirt::lsem.bootstrap()` function can be found in Listing 5. A slight misfit is detected in fit statistics RMSEA and SRMR. The CFI and TLI fit statistics are not indicative of the incorrect invariance assumption of factor loadings.

Figure 1 displays parameter curves for the two factor variances (i.e., `FX~~FX` and `FY~~FY`) and the factor correlation (i.e., `std_FX~~FY`) for the illustrative dataset *DATA1*. From Listing 5, we see that the variance of `FX` had an average of 0.396 with significant parameter variation ($SD_{bc} = 0.083$, $p < 0.001$), and `FY` had an average of 0.473 with significant parameter variation ($SD_{bc} = 0.111$, $p < 0.001$). Moreover, the factor correlation had an average of 0.584 and also showed a significant parameter variation ($SD_{bc} = 0.059$, $p = 0.003$).

Figure 2 displays parameter curves for the two factor variances and the factor correlation for the illustrative dataset *DATA3*, which had no simulated parameter variation in these parameters. By comparing Figures 1 and 2, it is evident that there is negligible parameter variation for the dataset *DATA3* compared to the dataset *DATA1*.

The parameter curves for DIF effects for factor loadings for datasets *DATA1* and *DATA2* are displayed in Figures 3 and 4, respectively. For *DATA1*, factor loadings were simulated as noninvariant, while they were assumed invariant across age for *DATA2*. This fact is visible when comparing Figures 3 and 4.

It can be seen from Listing 6 that DIF effects for factor loadings `X1` ($SD_{bc} = 0.024$, $p = 0.020$), `X3` ($SD_{bc} = 0.038$, $p = 0.002$), `Y1` ($SD_{bc} = 0.024$, $p = 0.022$), and `Y2` ($SD_{bc} = 0.030$, $p = 0.001$) had significant parameter variation for dataset *DATA1*, while they were not significant for loadings of `X2` and `Y3`.

Listing 6. Illustrative datasets: Part of the output of `sirt::lsem.bootstrap()` for the illustrative dataset *DATA1*.

```
Global Fit Statistics for Joint Estimation

   stat value value_bc    se
1 rmsea 0.017   -0.017 0.007
2   cfi 0.999    1.008 0.003
3   tli 0.998    1.011 0.004
4   gfi 0.988    0.996 0.003
5  srmr 0.028    0.016 0.003

Parameter Estimate Summary

            par parindex      M    SD SD_bc SD_se  SD_t  SD_p   MAD    Min    Max
1       FX=~X1         1  1.000 0.000 0.000 0.000 0.000 0.500 0.000  1.000  1.000
2       FX=~X2         2  1.099 0.000 0.000 0.000 0.000 0.500 0.000  1.099  1.099
3       FX=~X3         3  0.984 0.000 0.000 0.000 0.000 0.500 0.000  0.984  0.984
4       FY=~Y1         4  1.000 0.000 0.000 0.000 0.000 0.500 0.000  1.000  1.000
5       FY=~Y2         5  0.866 0.000 0.000 0.000 0.000 0.500 0.000  0.866  0.866
6       FY=~Y3         6  0.924 0.000 0.000 0.000 0.000 0.500 0.000  0.924  0.924
7       FX~~FX         7  0.396 0.086 0.083 0.011 7.635 0.000 0.064  0.148  0.478
8       FY~~FY         8  0.473 0.115 0.111 0.021 5.182 0.000 0.089  0.313  0.760
9       FX~~FY         9  0.258 0.077 0.073 0.011 6.538 0.000 0.059  0.083  0.400
[...]
32 std__FX~~FY        32  0.584 0.072 0.059 0.022 2.736 0.003 0.054  0.387  0.664
[...]
47 dif__FX=~X1        47  0.996 0.040 0.024 0.012 2.052 0.020 0.032  0.943  1.089
48 dif__FX=~X2        48  1.094 0.025 0.000 0.012 0.000 0.500 0.018  1.023  1.127
49 dif__FX=~X3        49  0.995 0.043 0.038 0.013 2.812 0.002 0.029  0.946  1.125
50 dif__FY=~Y1        50  1.011 0.038 0.024 0.012 2.013 0.022 0.031  0.947  1.076
51 dif__FY=~Y2        51  0.860 0.038 0.030 0.010 3.052 0.001 0.032  0.794  0.917
52 dif__FY=~Y3        52  0.925 0.024 0.000 0.011 0.000 0.500 0.021  0.882  0.970
```

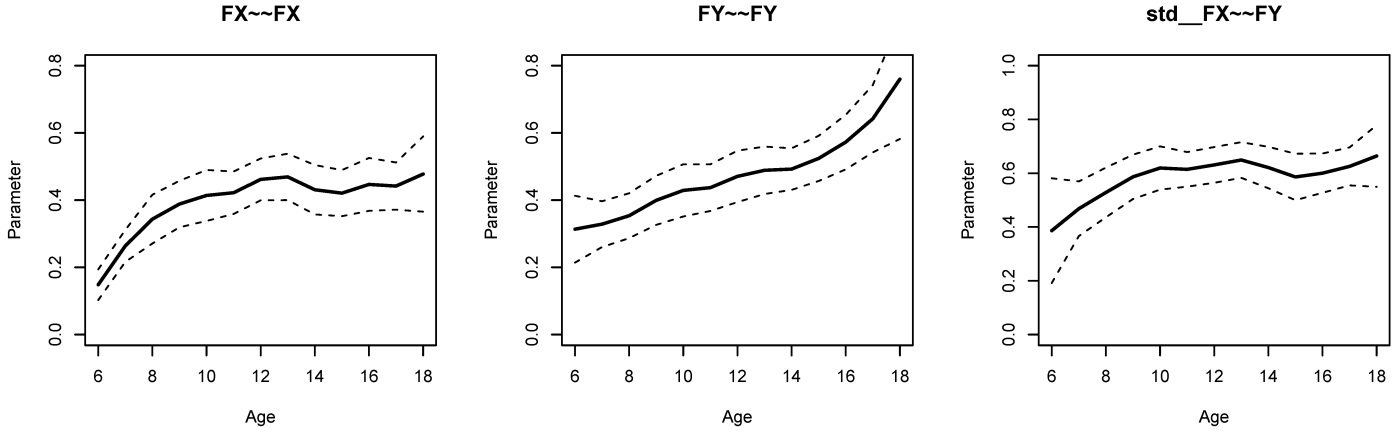

Figure 1. Illustrative datasets: Parameter curves for variances of the two factors (i.e., `FX~~FX` and `FX~~FX`) and the correlation of the two factors (`std__FX~~FY`) for the illustrative dataset *DATA1*.

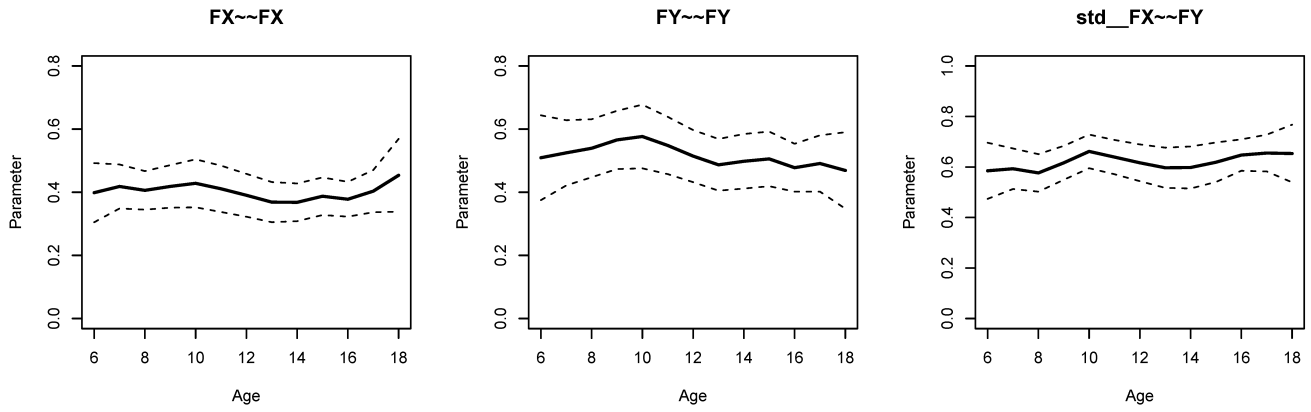

Figure 2. Illustrative datasets: Parameter curves for variances of the two factors (i.e., `FX`∼∼`FX` and `FX`∼∼`FX`) and the correlation of the two factors (`std__FX`∼∼`FY`) for the illustrative dataset *DATA3*.

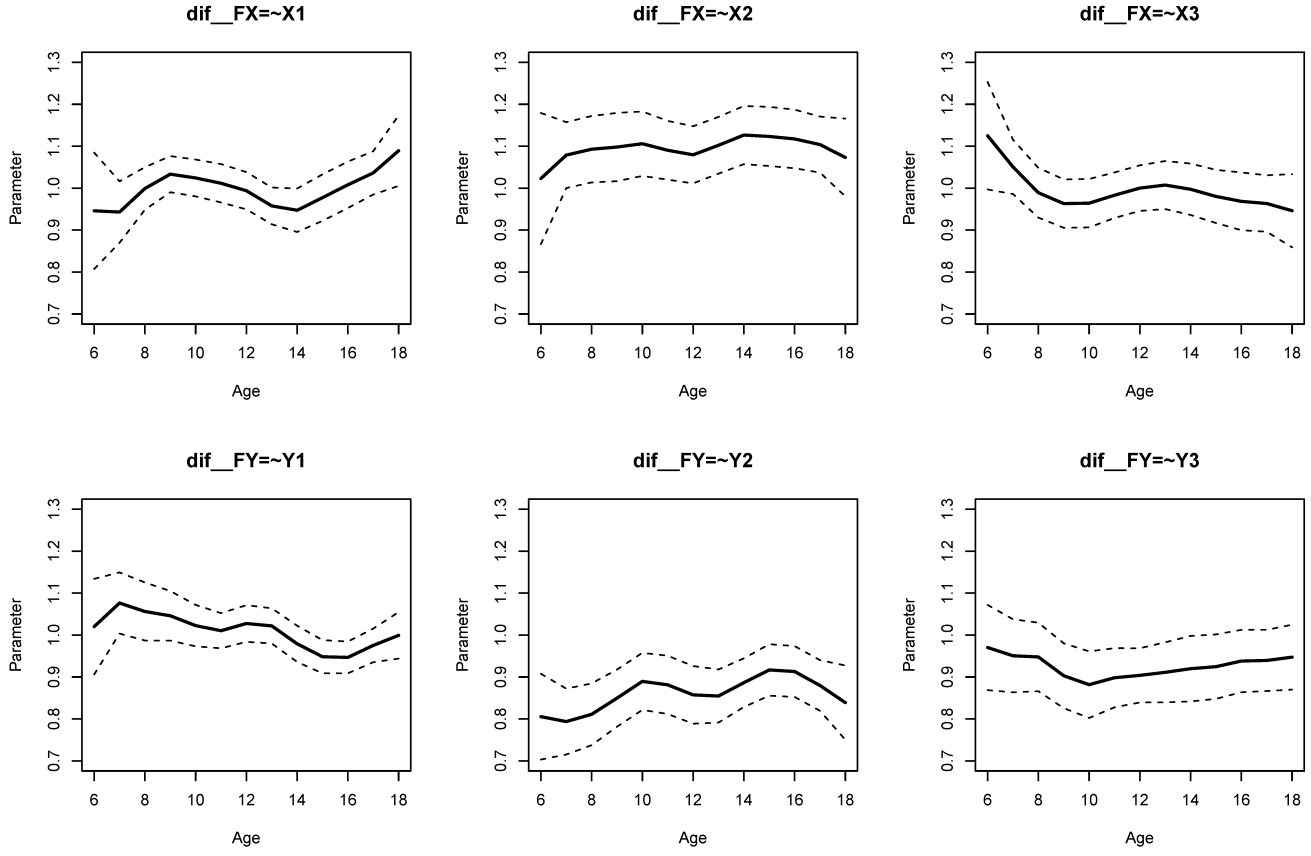

Figure 3. Illustrative datasets: Parameter curves for DIF effects of factor loadings for the illustrative dataset *DATA1*.

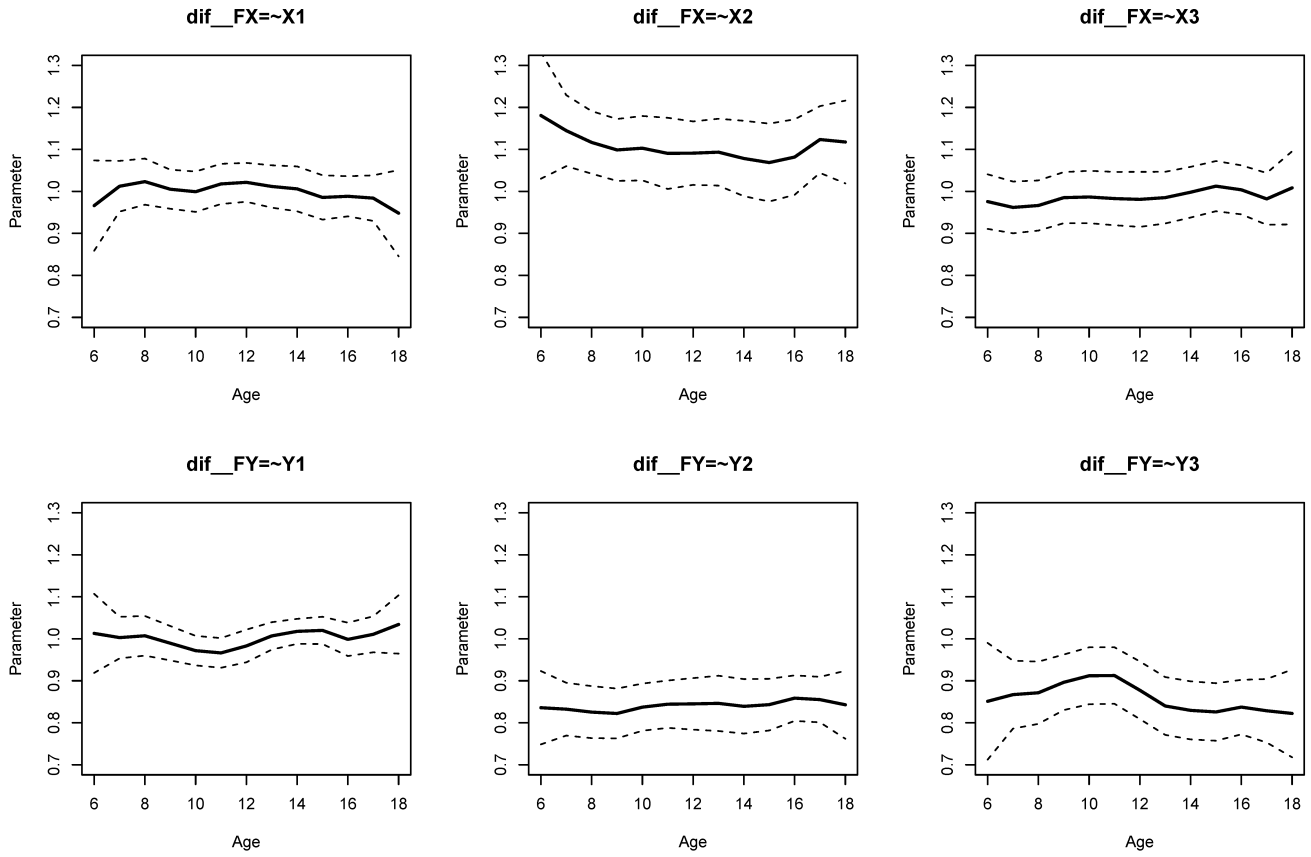

Figure 4. Illustrative datasets: Parameter curves for DIF effects of factor loadings for the illustrative dataset *DATA2*.

Finally, part of the R output of `sirt::lsem.bootstrap()` for dataset *DATA3* is displayed in Listing 7. In accordance with the data-generating model, both factor variances, the factor correlation, and the DIF effects for factor loadings did not show significant parameter variation across age.

Listing 7. Illustrative datasets: Part of the output of `sirt::lsem.bootstrap()` for the illustrative dataset *DATA3*.

```
Parameter Estimate Summary

            par parindex      M    SD SD_bc SD_se  SD_t  SD_p   MAD    Min   Max
[...]
7        FX~~FX        7  0.402 0.023 0.000 0.011 0.000 0.500 0.018  0.368 0.454
8        FY~~FY        8  0.517 0.032 0.000 0.014 0.000 0.500 0.027  0.469 0.577
9        FX~~FY        9  0.282 0.021 0.000 0.009 0.000 0.500 0.017  0.253 0.329
[...]
32 std__FX~~FY       32  0.619 0.028 0.000 0.012 0.000 0.500 0.024  0.577 0.662
[...]
47 dif__FX=~X1       47  1.000 0.019 0.000 0.011 0.000 0.500 0.016  0.948 1.022
48 dif__FX=~X2       48  1.106 0.027 0.000 0.012 0.000 0.500 0.021  1.069 1.175
49 dif__FX=~X3       49  0.987 0.014 0.000 0.006 0.000 0.500 0.011  0.965 1.013
50 dif__FY=~Y1       50  1.001 0.019 0.000 0.009 0.000 0.500 0.015  0.967 1.037
51 dif__FY=~Y2       51  0.842 0.012 0.000 0.007 0.000 0.500 0.010  0.822 0.863
52 dif__FY=~Y3       52  0.862 0.034 0.014 0.012 1.161 0.123 0.030  0.805 0.914
```

Note that a researcher will only have one dataset available for analysis. This section shows that LSEM model parameter output and figures are able to distinguish between situations of noninvariant and invariant model parameters. The standard deviation of a model parameter quantifies the variability of a model parameter across the values of the moderator.

For identification and interpretation reasons, it is useful to specify LSEM models with (some) invariant factor loadings. DIF effects reported in the LSEM output provide a post hoc assessment of the variability of parameter curves across the moderator values if parameter invariance was specified in the LSEM.

6. Simulation Study 1: Bias and RMSE

6.1. Method

In Simulation Study 1, the bias and the root mean square error (RMSE) of LSEM estimates of parameter curves were investigated. A one-factor model for three indicators, `X1`, `X2`, and `X3`, with a latent factor variable `FX` was specified. The data-generating model coincided with those from the illustrative datasets presented in Section 5. In contrast to Section 5, we only used the first three observed variables and considered a one-factor instead of a two-factor model in Simulation Study 1.

The population parameters can be found in the directory "*POPPARS*" at https://osf.io/puaz9/?view_only=63ffb2fd30f5400e89c59d03366bf793 (accessed on 3 June 2023). In this simulation, sample sizes N were chosen as 250, 500, 1000, 2000, and 4000. Instead of simulating data, random samples without replacement of sample size N were drawn from population datasets *DATA1* (noninvariant factor loadings, noninvariant factor variances and correlations), resulting in the data-generating model (DGM) DGM1, *DATA2* (invariant factor loadings, noninvariant factor variances and correlations) resulting in DGM2, and *DATA3* (invariant factor loadings, invariant factor variances and correlations), resulting in DGM3. The population datasets that included 130,000 subjects each can be found in the directory "*POPDATA*" at https://osf.io/puaz9/?view_only=63ffb2fd30f5400e89c59d03366bf793 (accessed on 3 June 2023).

Joint LSEM estimation was carried out using invariant item loadings and bandwidth factor $h = 1.1$, 2, and 3, where the bandwidth bw was defined as $bw = hN^{-1/5}\hat{\sigma}_A$. The Gaussian kernel function was used. We also compared the two choices of computing conditional covariances with local smoothing (SM; see (17)) and the weighting approach (16) (no smoothing; NSM). Moreover, we applied LSEM with a quadratic parameter constraint ("quad") using a bandwidth factor $h = 1.1$. A grid of 13 focal points was chosen as $6, 7, \ldots, 18$.

We investigated the accuracy of the estimated parameter curves of the variance of the latent factor F_X, the invariant factor loading of the indicator X_2, and the DIF effect for factor loading of X_2. Parameter accuracy was assessed by summarizing bias and RMSE of estimated parameter curves across the different age values. The bias of a parameter $\theta(a_t)$ at a focal point a_t is given by:

$$Bias(\hat{\theta}(a_t)) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r(a_t) - \theta(a_t)\right) , \tag{33}$$

where $\hat{\theta}_r(a_t)$ is the parameter estimate of $\theta(a_t)$ in the rth replication. The weighted absolute bias can then be defined as:

$$wBias(\hat{\theta}) = \sum_{t=1}^{T} f(a_t)|Bias(\hat{\theta}(a_t))| , \tag{34}$$

where $f(a_t)$ denotes the proportion of values of the moderator variable that equal a_t. The weighted root mean square error (weighted RMSE) is defined as:

$$wRMSE(\hat{\theta}) = \sum_{t=1}^{T} f(a_t) \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}_r(a_t) - \theta(a_t)\right)^2}, \tag{35}$$

which is a weighted point-wise RMSE summary statistic.

In total, 2500 replications (i.e., 2500 datasets were generated and analyzed in each condition of the simulation) were conducted. We used the R (R Core Team 2023) software for the entire analysis of the simulation and the sirt (Robitzsch 2023b) package for LSEM estimation.

6.2. Results

In Table 1, weighted absolute bias and weighted RMSE for the factor variance, the invariant factor loading of X_2, and the DIF effect of factor loading of X_2 are presented.

It turned out that all three model parameters resulted in unbiased estimation for moderate or large sample sizes. For DGM1 or DGM2, the quadratic parameter constraint introduced some misspecification, which led to slight biases. Moreover, using the local quadratic smoothing approach SM for estimating conditional covariances instead of the weighted approach NM (e.g., no smoothing) resulted in a small error bias. Finally, biases increased with increasing the bandwidth factor h.

Notably, using local smoothing SM for conditional covariances added variability in terms of RMSE compared to NM. Regarding RMSE, one could conclude that $h = 2$ seems preferable to $h = 1.1$ or $h = 3$ (see also Hildebrandt et al. 2016).

Overall, the findings of Simulation Study 1 demonstrated that joint LSEM estimation resulted in approximately unbiased parameter estimates. The decrease in RMSE values for increasing sample sizes also indicated that parameter estimates are consistent. Notably, the recommendation of using the bandwidth factor $h = 2$ in pointwise LSEM (Hildebrandt et al. 2016) also transfers to the joint LSEM estimation method.

Table 1. Simulation Study 1: Weighted absolute bias and weighted root mean square error (RMSE) for the parameter curve $\theta(a)$ for different model parameters as a function of sample size N and three data-generating models DGM1, DGM2 and DGM3.

		Weighted Absolute Bias							**Weighted RMSE**						
		$h = 1.1$		$h = 2$		$h = 3$			$h = 1.1$		$h = 2$		$h = 3$		
DGM	N	**SM**	**NSM**	**SM**	**NSM**	**SM**	**NSM**	**Quad**	**SM**	**NSM**	**SM**	**NSM**	**SM**	**NSM**	**Quad**
		Variance of latent factor F													
	250	0.022	0.022	0.021	0.028	0.025	0.039	0.023	0.088	0.083	0.078	0.076	0.077	0.079	0.076
	500	0.013	0.014	0.015	0.022	0.020	0.033	0.016	0.064	0.061	0.057	0.056	0.056	0.060	0.054
1	1000	0.007	0.009	0.010	0.018	0.016	0.029	0.011	0.048	0.046	0.042	0.043	0.042	0.048	0.039
	2000	0.006	0.008	0.008	0.015	0.013	0.025	0.010	0.035	0.034	0.031	0.033	0.032	0.038	0.029
	4000	0.003	0.005	0.005	0.012	0.010	0.021	0.008	0.026	0.026	0.023	0.025	0.023	0.030	0.021
	250	0.023	0.019	0.021	0.022	0.022	0.030	0.018	0.086	0.080	0.076	0.070	0.073	0.070	0.073
	500	0.013	0.011	0.014	0.017	0.017	0.026	0.011	0.063	0.059	0.055	0.053	0.054	0.054	0.052
2	1000	0.008	0.008	0.010	0.015	0.014	0.023	0.008	0.046	0.044	0.040	0.040	0.040	0.042	0.038
	2000	0.005	0.006	0.007	0.012	0.012	0.020	0.006	0.034	0.033	0.030	0.030	0.029	0.032	0.027
	4000	0.003	0.004	0.005	0.009	0.009	0.017	0.005	0.025	0.025	0.022	0.023	0.022	0.026	0.020
	250	0.018	0.017	0.012	0.010	0.009	0.007	0.017	0.087	0.080	0.076	0.067	0.071	0.061	0.073
	500	0.010	0.009	0.006	0.005	0.004	0.004	0.009	0.063	0.059	0.055	0.049	0.051	0.044	0.052
3	1000	0.006	0.006	0.004	0.004	0.003	0.003	0.006	0.047	0.044	0.040	0.036	0.037	0.033	0.038
	2000	0.002	0.002	0.001	0.001	0.001	0.001	0.002	0.034	0.032	0.029	0.026	0.026	0.023	0.026
	4000	0.002	0.002	0.001	0.001	0.001	0.001	0.002	0.025	0.024	0.021	0.020	0.019	0.017	0.019

Table 1. *Cont.*

		Weighted Absolute Bias							Weighted RMSE						
		$h = 1.1$		$h = 2$		$h = 3$			$h = 1.1$		$h = 2$		$h = 3$		
DGM	N	**SM**	**NSM**	**SM**	**NSM**	**SM**	**NSM**	**Quad**	**SM**	**NSM**	**SM**	**NSM**	**SM**	**NSM**	**Quad**
		Invariant factor loading of X_2													
	250	0.003	0.005	0.006	0.004	0.008	0.004	0.005	0.089	0.090	0.089	0.089	0.090	0.088	0.089
	500	0.001	0.001	0.000	0.000	0.002	0.000	0.001	0.063	0.063	0.062	0.063	0.063	0.062	0.063
1	1000	0.001	0.000	0.001	0.000	0.001	0.000	0.000	0.043	0.043	0.043	0.043	0.044	0.043	0.043
	2000	0.001	0.002	0.000	0.002	0.002	0.001	0.001	0.031	0.031	0.030	0.031	0.031	0.031	0.031
	4000	0.000	0.001	0.001	0.001	0.000	0.000	0.000	0.022	0.022	0.022	0.022	0.022	0.022	0.022
	250	0.005	0.005	0.004	0.004	0.004	0.004	0.005	0.091	0.091	0.091	0.090	0.091	0.090	0.091
	500	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.062	0.062	0.062	0.062	0.062	0.061	0.062
2	1000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.045	0.045	0.045	0.045	0.045	0.045	0.045
	2000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.031	0.031	0.031	0.031	0.031	0.031	0.031
	4000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.022	0.022	0.022	0.022	0.022	0.022	0.022
	250	0.003	0.003	0.002	0.003	0.002	0.003	0.003	0.090	0.090	0.090	0.090	0.090	0.090	0.090
	500	0.003	0.003	0.003	0.003	0.003	0.002	0.003	0.064	0.064	0.063	0.064	0.064	0.064	0.064
3	1000	0.000	0.001	0.000	0.001	0.000	0.001	0.001	0.043	0.043	0.043	0.043	0.043	0.044	0.043
	2000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.032	0.032	0.032	0.032	0.032	0.032	0.032
	4000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.022	0.022	0.022	0.022	0.022	0.022	0.022
		DIF for factor loading of X_2													
	250	0.007	0.007	0.010	0.012	0.013	0.015	0.013	0.123	0.109	0.111	0.098	0.105	0.094	0.123
	500	0.004	0.006	0.008	0.011	0.012	0.014	0.013	0.085	0.079	0.078	0.071	0.075	0.068	0.092
1	1000	0.003	0.004	0.006	0.009	0.010	0.013	0.012	0.061	0.057	0.055	0.051	0.053	0.049	0.069
	2000	0.002	0.004	0.005	0.008	0.008	0.012	0.012	0.044	0.041	0.040	0.037	0.038	0.036	0.053
	4000	0.002	0.003	0.004	0.007	0.007	0.010	0.012	0.032	0.031	0.029	0.027	0.028	0.027	0.041
	250	0.005	0.005	0.005	0.004	0.004	0.004	0.009	0.121	0.110	0.109	0.098	0.105	0.094	0.122
	500	0.002	0.002	0.002	0.002	0.002	0.002	0.008	0.083	0.077	0.075	0.069	0.072	0.065	0.089
2	1000	0.001	0.001	0.001	0.001	0.001	0.001	0.008	0.061	0.058	0.055	0.051	0.053	0.048	0.068
	2000	0.001	0.001	0.001	0.001	0.001	0.001	0.007	0.043	0.041	0.039	0.036	0.037	0.034	0.051
	4000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.032	0.030	0.028	0.026	0.027	0.025	0.039
	250	0.003	0.003	0.003	0.003	0.002	0.003	0.002	0.118	0.109	0.106	0.098	0.102	0.094	0.120
	500	0.003	0.003	0.003	0.003	0.003	0.003	0.002	0.083	0.079	0.076	0.071	0.073	0.067	0.090
3	1000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.059	0.056	0.053	0.050	0.050	0.047	0.066
	2000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.044	0.042	0.039	0.037	0.037	0.035	0.051
	4000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.031	0.030	0.028	0.026	0.026	0.025	0.038

7. Simulation Study 2: Estimation of Variability of Model Parameters and Statistical Significance Tests

7.1. Method

In Simulation Study 2, the bias of standard deviation statistics for parameter variation and the properties of significance tests for parameter variation are investigated. The same three data-generating models DGM1, DGM2, and DGM3 as in Simulation Study 1 (see Section 6.1) were utilized.

The chosen sample sizes in this simulation were $N = 500$, 1000, 2000, and 4000. As in Simulation Study 1, samples of sample size N were drawn without replacement from population datasets *DATA1*, *DATA2*, and *DATA3* for DGM1, DGM2, and DGM3, respectively. The population datasets that included 130,000 subjects can be found in the directory "*POPDATA*" at https://osf.io/puaz9/?view_only=63ffb2fd30f5400e89c59d03366bf793 (accessed on 3 June 2023).

As in Simulation Study 1, a one-factor model with indicators `X1`, `X2`, and `X3` was specified. Throughout all simulation conditions, a bandwidth factor of $h = 2$ was chosen. The bias of the two standard deviation estimators $\widehat{\mathrm{SD}}_{\theta(a)}$ and $\widehat{\mathrm{SD}}_{\theta(a),\mathrm{bc}}$ defined in (28) and (29) was assessed. Significance testing for parameter variation was based on the standard deviation (see (30)), which uses a normal distribution approximation and the Wald test

(see (31)), which uses a chi-square distribution as a null distribution. Statistical significance tests were performed with significance levels of 0.05 and 0.01. The bias of the standard deviation variability statistics and significance tests of parameter variation was computed for the variance of the latent factor, the three DIF effects of the factor loadings, and the three residual variances.

In total, 2500 replications were conducted in all simulation conditions. The R software (R Core Team 2023) was used for analyzing this simulation study, and the R package sirt (Robitzsch 2023b) was employed for LSEM estimation and significance testing.

7.2. Results

In Table 2, the bias of raw and bias-corrected ("bc") estimates of the standard deviation variability measure $SD_{\theta(a)}$ are presented. In DGM1, all parameters have nonvanishing $SD_{\theta(a)}$ values for the population dataset *DATA1*. In this case, the raw SD estimate showed some slight positive bias for sample sizes $N = 500$ and 1000. The bias-corrected estimates were generally negatively biased, although the biases were not very large. In DGM2, only the variance of the latent factor F had a true parameter variation larger than 0. In this situation, raw estimates were approximately unbiased, while the bias-corrected estimates were negatively biased. If there was no true parameter variation, such as for DIF effects or residual variances in DGM2 or all parameters in DGM3, the bias-corrected estimates were less biased than the raw standard deviation estimate.

Table 2. Simulation Study 2: Bias of raw and bias-corrected estimators of the standard deviation $SD_{\theta(a)}$ for the parameter curve $\theta(a)$ for different model parameters as a function of sample size N and three data-generating models DGM1, DGM2 and DGM3.

	DGM 1			**DGM 2**			**DGM 3**		
	$\mathbf{SD_{\theta(a)}}$			$\mathbf{SD_{\theta(a)}}$			$\mathbf{SD_{\theta(a)}}$		
N	**true**	**raw**	**bc**	**true**	**raw**	**bc**	**true**	**raw**	**bc**
	Variance of latent factor F								
500	0.081	−0.002	−0.012	0.054	0.002	−0.013	0	0.035	0.013
1000	0.081	−0.003	−0.009	0.054	−0.001	−0.009	0	0.027	0.010
2000	0.081	−0.003	−0.006	0.054	−0.002	−0.006	0	0.020	0.007
4000	0.081	−0.003	−0.005	0.054	−0.002	−0.004	0	0.015	0.005
	DIF for factor loading of X_1								
500	0.047	0.007	−0.021	0	0.041	0.013	0	0.039	0.013
1000	0.047	0.001	−0.017	0	0.031	0.010	0	0.029	0.010
2000	0.047	−0.002	−0.012	0	0.023	0.007	0	0.022	0.007
4000	0.047	−0.002	−0.008	0	0.017	0.005	0	0.017	0.005
	DIF for factor loading of X_2								
500	0.021	0.021	−0.007	0	0.038	0.012	0	0.036	0.012
1000	0.021	0.013	−0.007	0	0.029	0.009	0	0.027	0.009
2000	0.021	0.006	−0.008	0	0.022	0.007	0	0.021	0.007
4000	0.021	0.003	−0.007	0	0.017	0.005	0	0.016	0.005
	DIF for factor loading of X_3								
500	0.022	0.009	−0.008	0	0.022	0.006	0	0.021	0.006
1000	0.022	0.005	−0.006	0	0.017	0.005	0	0.016	0.005
2000	0.022	0.002	−0.004	0	0.013	0.004	0	0.012	0.004
4000	0.022	0.001	−0.002	0	0.009	0.003	0	0.009	0.003
	Residual variance of X_1								
500	0.012	0.014	−0.001	0	0.024	0.009	0	0.024	0.009
1000	0.012	0.009	−0.003	0	0.018	0.006	0	0.018	0.006
2000	0.012	0.005	−0.003	0	0.014	0.005	0	0.014	0.005
4000	0.012	0.003	−0.003	0	0.011	0.003	0	0.011	0.003

Table 2. *Cont.*

	DGM 1			DGM 2			DGM 3		
	$SD_{\theta(a)}$			$SD_{\theta(a)}$			$SD_{\theta(a)}$		
N	**true**	**raw**	**bc**	**true**	**raw**	**bc**	**true**	**raw**	**bc**
	Residual variance of X_2								
500	0.007	0.020	0.003	0	0.027	0.010	0	0.027	0.010
1000	0.007	0.014	0.001	0	0.021	0.007	0	0.020	0.007
2000	0.007	0.010	−0.001	0	0.016	0.005	0	0.016	0.005
4000	0.007	0.006	−0.002	0	0.012	0.004	0	0.012	0.004
	Residual variance of X_3								
500	0.011	0.009	−0.002	0	0.018	0.007	0	0.018	0.007
1000	0.011	0.006	−0.002	0	0.014	0.005	0	0.014	0.005
2000	0.011	0.003	−0.003	0	0.010	0.003	0	0.011	0.004
4000	0.011	0.002	−0.002	0	0.008	0.002	0	0.008	0.003

true = true value of $SD_{\theta(a)}$ in infinite sample size (i.e., at the population level); raw = raw estimate $\widehat{SD}_{\theta(a)}$ of $SD_{\theta(a)}$ (see Equation (28)); bc = bias-corrected estimate $\widehat{SD}_{\theta(a),bc}$ of $SD_{\theta(a)}$ (see Equation (29)).

Overall, one could say that for smaller values of true variability, the positive bias in the raw SD estimate was larger than the underestimation of the bias-corrected SD estimate. An improved SD statistic might be obtained by computing some weighted average of the raw and the bias-corrected estimate.

Table 3 presents type I error and power rates for the different LSEM model parameters. Significance testing based on the SD statistics had inflated type I error rates. If the nominal level was chosen as 1%, the empirical error rate was about 5%. Moreover, the Wald statistic had type I error rates lower than the nominal level in many simulation conditions. Nevertheless, significance testing based on the standard deviation has substantially more statistical power. If a target nominal significance level for the SD test statistic were 5%, it is advised to use a significance level of 0.01.

Table 3. Simulation Study 2: Type I and power rates for the significance test for variability in a parameter curve $\theta(a)$ for the two test statistics based on $SD_{\theta(a)}$ (SD) and the Wald test (WA) as a function of sample size N and three data-generating models DGM1, DGM2 and DGM3.

	DGM1				DGM2				DGM3			
N	**SD5**	**WA5**	**SD1**	**WA1**	**SD5**	**WA5**	**SD1**	**WA1**	**SD5**	**WA5**	**SD1**	**WA1**
	Variance of latent factor F											
500	92.4	46.9	79.8	29.6	66.0	17.8	44.2	8.6	16.3	1.6	4.9	0.5
1000	99.7	88.1	98.5	75.4	90.0	45.7	76.1	26.5	17.4	2.8	5.9	0.8
2000	100.0	99.8	100.0	99.1	99.1	83.0	97.0	65.8	17.7	3.6	5.8	0.9
4000	100.0	100.0	100.0	100.0	100.0	99.5	100.0	98.1	18.3	6.4	7.7	2.0
	DIF for factor loading of X_1											
500	23.7	1.3	8.2	0.4	12.9	0.3	3.4	0.1	13.2	0.4	3.4	0.0
1000	46.1	6.0	22.8	1.2	14.8	0.8	5.2	0.2	14.7	0.5	5.1	0.1
2000	75.2	25.2	52.1	10.7	14.4	1.6	4.9	0.3	15.9	1.7	5.5	0.2
4000	96.2	70.6	89.6	48.7	17.5	4.0	6.1	1.1	17.9	4.1	7.1	1.1
	DIF for factor loading of X_2											
500	12.3	0.4	3.3	0.1	12.4	0.3	3.5	0.0	12.3	0.3	3.3	0.0
1000	21.7	2.2	8.6	0.5	13.1	0.7	4.1	0.1	13.9	0.8	4.7	0.1
2000	31.2	5.3	14.9	1.5	16.7	1.8	5.8	0.4	16.3	1.4	5.3	0.4
4000	52.4	19.1	32.1	8.2	16.2	3.6	6.9	0.7	18.1	4.2	7.1	1.3

Table 3. *Cont.*

	DGM1				DGM2				DGM3			
N	**SD5**	**WA5**	**SD1**	**WA1**	**SD5**	**WA5**	**SD1**	**WA1**	**SD5**	**WA5**	**SD1**	**WA1**
	DIF for factor loading of X_3											
500	18.4	0.4	5.4	0.1	7.5	0.2	1.6	0.0	8.3	0.1	1.8	0.0
1000	38.6	2.0	16.5	0.4	10.5	0.4	2.5	0.0	12.1	0.3	2.9	0.0
2000	66.1	13.7	42.5	4.6	14.2	0.8	4.6	0.0	13.8	1.0	4.6	0.2
4000	92.3	52.0	81.0	28.8	16.1	2.3	5.8	0.5	17.7	2.8	6.2	0.4
	Residual variance of X_1											
500	20.7	2.3	7.5	0.9	18.2	2.2	6.8	0.5	16.6	2.0	6.2	0.7
1000	25.3	4.0	10.4	1.6	16.6	2.8	5.6	0.9	17.8	3.1	6.5	0.8
2000	34.7	8.4	17.4	2.6	18.3	4.4	6.4	1.2	18.0	4.3	7.1	1.0
4000	49.1	17.7	29.1	7.1	17.3	5.7	7.0	1.5	19.0	6.0	7.2	1.6
	Residual variance of X_2											
500	16.9	1.9	5.9	0.4	17.7	1.9	6.4	0.4	17.1	1.8	6.3	0.6
1000	18.7	2.5	7.0	0.7	17.6	2.8	6.2	0.7	17.8	2.8	5.7	0.9
2000	22.1	4.8	8.4	1.3	18.5	4.4	7.1	1.3	16.7	3.6	5.9	1.0
4000	29.6	8.8	13.3	3.0	18.0	5.9	7.4	1.6	19.0	6.5	7.1	1.6
	Residual variance of X_3											
500	25.4	2.9	11.2	0.8	16.7	1.4	5.6	0.5	17.7	2.0	6.1	0.5
1000	34.1	5.5	15.9	1.5	17.9	2.7	6.2	0.7	18.1	2.7	6.4	0.8
2000	47.3	12.0	26.6	4.8	17.2	3.3	6.4	0.8	18.3	4.0	6.7	1.1
4000	68.6	26.3	47.4	12.7	17.9	5.3	7.0	1.4	19.2	6.5	7.7	1.7

Note. SD5 = test statistic based on bias-corrected $SD_{\theta(a)}$ estimate at 5% confidence level; WA5 = Wald test statistic at 5% confidence level; SD1 = test statistic based on bias-corrected $SD_{\theta(a)}$ estimate at 1% confidence level; WA1 = Wald test statistic at 1% confidence level; Cells with yellow-gray colored background correspond to type I error rates, while cells with white background color correspond to power rates.

8. Empirical Example: A Reanalysis of SON-R

8.1. Data

According to the age differentiation hypothesis, cognitive abilities become more differentiated with increasing age during childhood. Hülür et al. (2011) used data from the German standardization of the SON-R $2^1/_2-7$ intelligence test to examine age-related differentiation of cognitive abilities from age $2^1/_2$ to age 7. The SON-R $2^1/_2-7$ intelligence test is a nonverbal intelligence test for children and consists of six indicators (i.e., six subtests). The SON-R $2^1/_2-7$ test contains two subscales measured by three indicators each. The performance subscale (with factor `Fp`) contains indicators mosaics (`p1`), puzzles (`p2`), and patterns (`p3`). The reasoning subscale (with factor `Fr`) contains the indicators categories (`r1`), analogies (`r2`), and situations (`r3`).

Unfortunately, the SON-R dataset is not publicly available, and the authors of this paper cannot publicly share the dataset on the internet. To replicate the LSEM analysis of this example, we generated a synthetic dataset of the SON-R $2^1/_2-7$ data based on the original dataset used in Hülür et al. (2011). The same sample size of $N = 1027$ children was simulated. In the synthetic data generation, we relied on a recently proposed method by Jiang et al. (2021) (see also Grund et al. 2022, Nowok et al. 2016 or Reiter 2023) that combines the distinct approaches of simulating a dataset based on a known distribution and the approach of adding to noise to original data to prevent data disclosure or person identification. The noisy versions of the original dataset were simulated with a reliability of 0.95 (Grund et al. 2022), and quadratic relations among variables were allowed. The data synthesis model was separately carried out in 18 groups of children (i.e., in 9 age groups for male and female children, respectively). The values of the age and gender variables were held fixed in the analysis meaning that these demographic variables had the same distribution in the synthetic data as in the original data. In total, 50.8% of the children in the sample was male. The synthetic data and syntax for synthetic data generation can be found

in the directory "*SON-R*" at https://osf.io/puaz9/?view_only=63ffb2fd30f5400e89c59d03366bf793 (accessed on 3 June 2023).

The indicator variables were linearly transformed such that the mean equaled zero and the standard deviation equaled one for children aged between 4.0 and 6.0 years. This is an arbitrary choice and only affects the scaling of the variables. The assessment of model parameter heterogeneity in the LSEM is independent of this choice. Alternatively, one might also standardize the indicator variables for children in the total sample with ages between 2.5 and 7.5 years.

A two-dimensional CFA model involving the performance and the reasoning factor was specified in an LSEM analysis. The mean structure remained unmodeled because the primary goal of this analysis was to investigate the age differentiation hypothesis. For model identification, the factor loadings were assumed as invariant across age, and the first loading of both scales (i.e., loadings of `p1` and `r1`) were fixed at one. In accordance with Hildebrandt et al. (2016) and the findings of Simulation Study 1, the bandwidth factor of $h = 2$ was chosen, resulting in a bandwidth $bw = 2N^{-1/5}\hat{\sigma}_A$, where $\hat{\sigma}_A = 1.23$ is the estimated standard deviation of the age variable. Because the LSEM model involved invariance constraints among parameters, a joint estimation approach was employed. For statistical inference and the test of parameter variation, $R = 200$ bootstrap samples were drawn. Replication syntax can also be found in the directory "*SON-R*" at https://osf.io/puaz9/?view_only=63ffb2fd30f5400e89c59d03366bf793 (accessed on 3 June 2023).

8.2. Results

Figure 5 displays the histogram of the age variable. The age of children ranged between 2.44 and 7.72 years, with a mean of 4.89 and a standard deviation of 1.23. The histogram indicated that the intended age range between 2.5 and 7 years of the SON-R $2^1/_2{-}7$ test was approximately uniformly distributed.

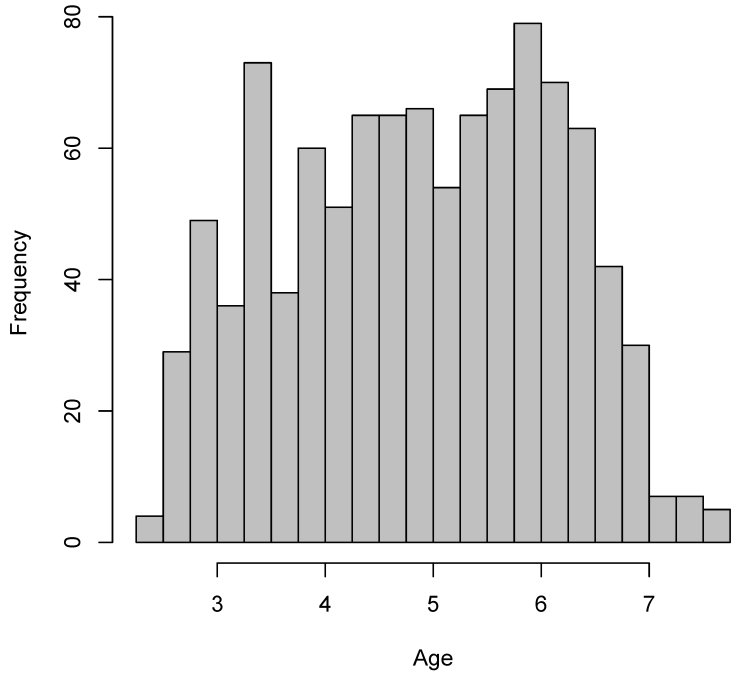

Figure 5. SON-R example: Histogram for moderator age.

The estimated LSEM model had an acceptable model fit regarding typical model fit effect sizes. The fit statistics without bias correction were RMSEA = 0.061, CFI = 0.952, TLI = 0.960, GFI = 0.963, and SRMR = 0.055.

In Listing 8, parts of the LSEM output of `lsem.bootstrap()` are displayed. According to the specified model, the parameter variation (i.e., `SD` and `SD_bc`) for factor loadings (i.e., `Fp=~p1`, ..., `F3=~r3`) was zero because the parameters were assumed invariant across age.

Listing 8. SON-R example: Part of the output of `lsem.bootstrap()` function.

```
Parameter Estimate Summary

            par parindex     M    SD SD_bc SD_se  SD_t  SD_p   MAD   Min   Max
1       Fp=~p1         1 1.000 0.000 0.000 0.000 0.000 0.500 0.000 1.000 1.000
2       Fp=~p2         2 0.854 0.000 0.000 0.000 0.000 0.500 0.000 0.854 0.854
3       Fp=~p3         3 0.803 0.000 0.000 0.000 0.000 0.500 0.000 0.803 0.803
4       Fr=~r1         4 1.000 0.000 0.000 0.000 0.000 0.500 0.000 1.000 1.000
5       Fr=~r2         5 1.041 0.000 0.000 0.000 0.000 0.500 0.000 1.041 1.041
6       Fr=~r3         6 1.160 0.000 0.000 0.000 0.000 0.500 0.000 1.160 1.160
7       Fp~~Fp         7 0.445 0.038 0.000 0.016 0.000 0.500 0.033 0.323 0.504
8       Fr~~Fr         8 0.232 0.019 0.000 0.013 0.000 0.500 0.017 0.181 0.260
9       Fp~~Fr         9 0.282 0.033 0.017 0.011 1.530 0.063 0.021 0.095 0.334
10      p1~~p1        10 0.347 0.082 0.076 0.020 3.777 0.000 0.057 0.032 0.532
11      p2~~p2        11 0.498 0.074 0.060 0.021 2.828 0.002 0.064 0.365 0.611
12      p3~~p3        12 0.411 0.046 0.030 0.015 2.057 0.020 0.031 0.122 0.573
13      r1~~r1        13 0.510 0.120 0.114 0.024 4.734 0.000 0.098 0.172 0.636
14      r2~~r2        14 0.510 0.110 0.103 0.024 4.206 0.000 0.096 0.333 0.675
15      r3~~r3        15 0.571 0.055 0.019 0.022 0.887 0.188 0.046 0.431 0.677
[...]
32 std__Fp~~Fr        32 0.878 0.071 0.019 0.041 0.480 0.316 0.041 0.395 0.949
[...]
47 dif__Fp=~p1        47 1.005 0.038 0.013 0.016 0.833 0.202 0.029 0.936 1.157
48 dif__Fp=~p2        48 0.868 0.089 0.077 0.029 2.662 0.004 0.077 0.400 0.985
49 dif__Fp=~p3        49 0.808 0.097 0.093 0.020 4.749 0.000 0.082 0.673 1.195
50 dif__Fr=~r1        50 0.994 0.068 0.000 0.043 0.000 0.500 0.054 0.770 1.233
51 dif__Fr=~r2        51 1.053 0.105 0.078 0.043 1.803 0.036 0.082 0.809 1.560
52 dif__Fr=~r3        52 1.159 0.126 0.092 0.045 2.056 0.020 0.107 0.567 1.338
```

The age differentiation hypothesis refers to the variances of the performance scale (i.e., `Fp~~Fp`), the variance of the reasoning scale (i.e., `Fr~~Fr`), and the correlation of both factors (i.e., `std__Fp~~Fr`). Figure 6 displays the parameter curves with confidence intervals for the two variances and the correlation. From the R output presented in Listing 8, it can be seen that the variances parameter curves did not show significant parameter variation, and the bias-corrected standard deviation estimate `SD_bc` was 0.000. The correlation between the performance and the reasoning scale was 0.878 on average, with a small bias-corrected standard deviation estimate of 0.019 that turned out to be nonsignificant ($p = 0.316$). Hence, there was no evidence for the age differentiation hypothesis in the SON-R dataset.

Figure 7 displays the parameter curves of the DIF effects of the factor loadings. The corresponding parameters for DIF effects can be found in lines 47 to 52 in Listing 8 (i.e., parameters `dif__Fp=~p1`, ..., `dif__Fr=~r3`). There was substantial parameter variation in terms of the bias-corrected standard deviation `SD_ bc` for the loadings of `p2` ($SD_{bc} = 0.077$, $p = 0.004$), `p3` ($SD_{bc} = 0.077$, $p < 0.001$), `r2` ($SD_{bc} = 0.078$, $p = 0.036$), and `r3` ($SD_{bc} = 0.092$, $p = 0.020$).

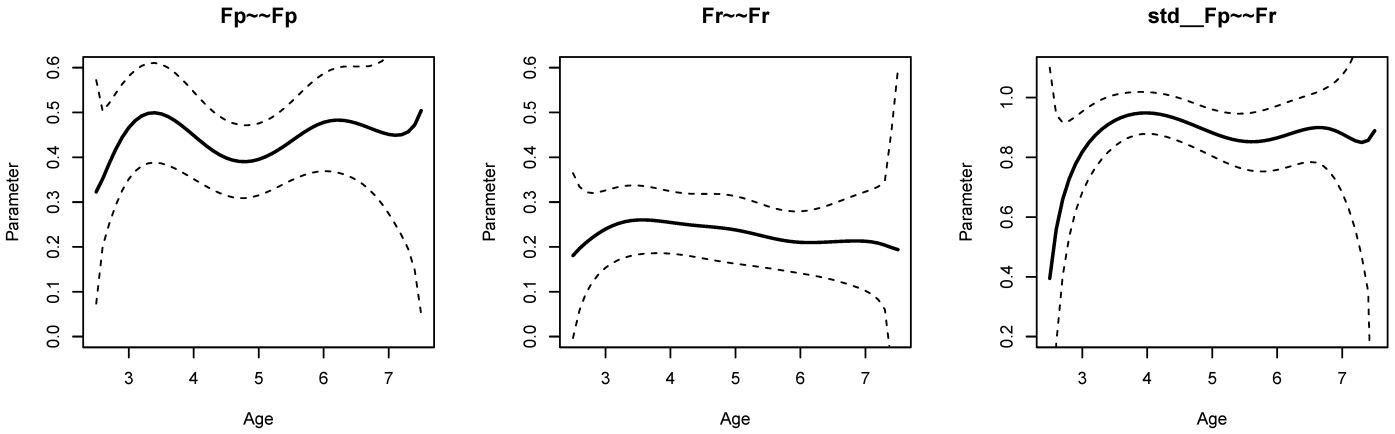

Figure 6. SON-R example: Parameter curves for variances of performance (`Fp`∼∼`Fp`) and reasoning (`Fr`∼∼`Fr`) and the correlation of performance and reasoning (`std__Fp`∼∼`Fr`).

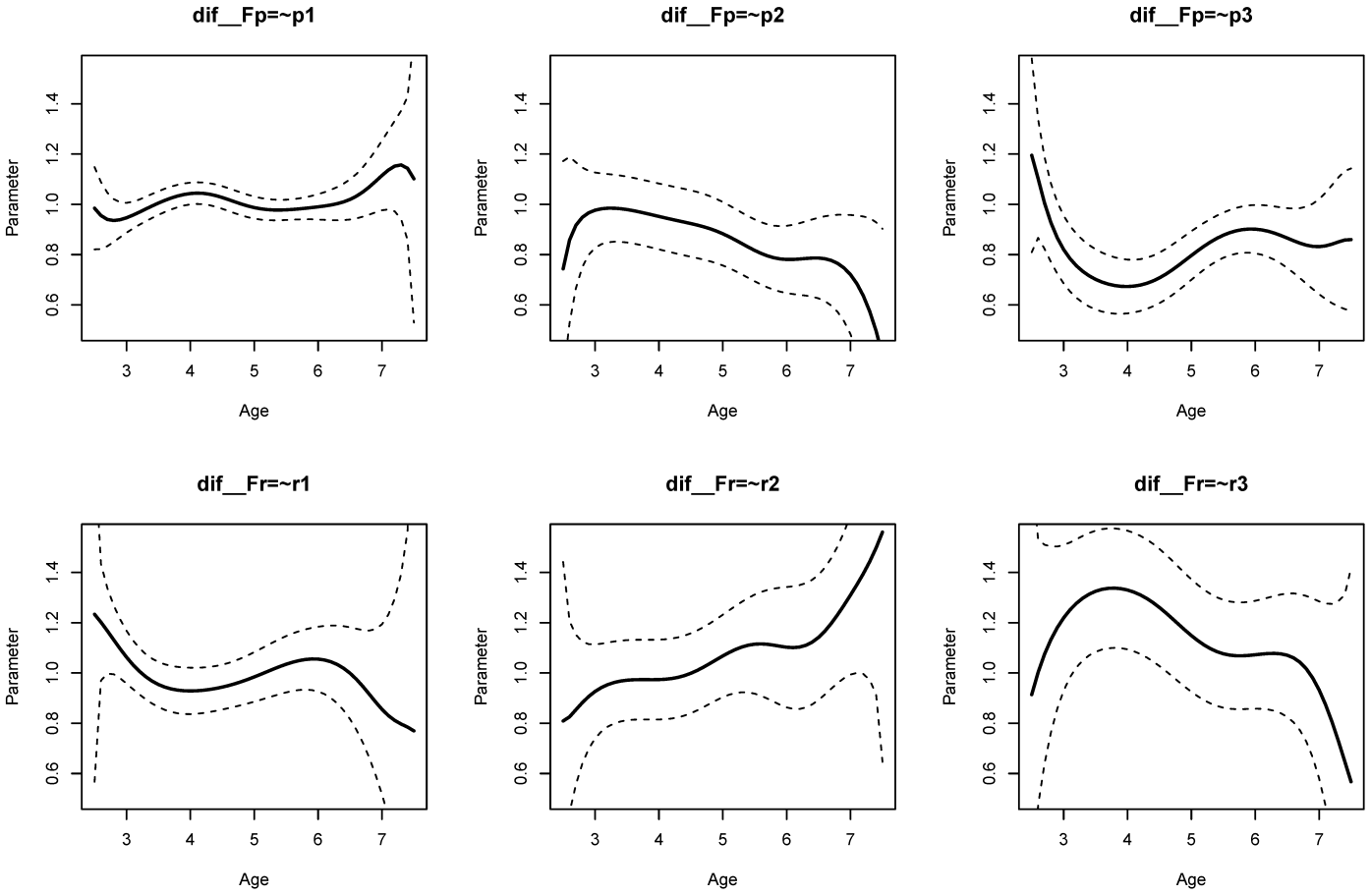

Figure 7. SON-R example: Parameter curves for DIF effects of factor loadings for performance (latent variable `Fp`) and reasoning (latent variable `Fr`).

Finally, residual variances are displayed in Figure 8. From the results from Listing 8, it is evident that residual variances of `p1` ($SD_{bc} = 0.076$, $p < 0.001$), `p2` ($SD_{bc} = 0.060$, $p = 0.002$), `r1` ($SD_{bc} = 0.114$, $p < 0.001$), and `r2` ($SD_{bc} = 0.103$, $p < 0.001$) were statistically significant at the 0.01 significance level.

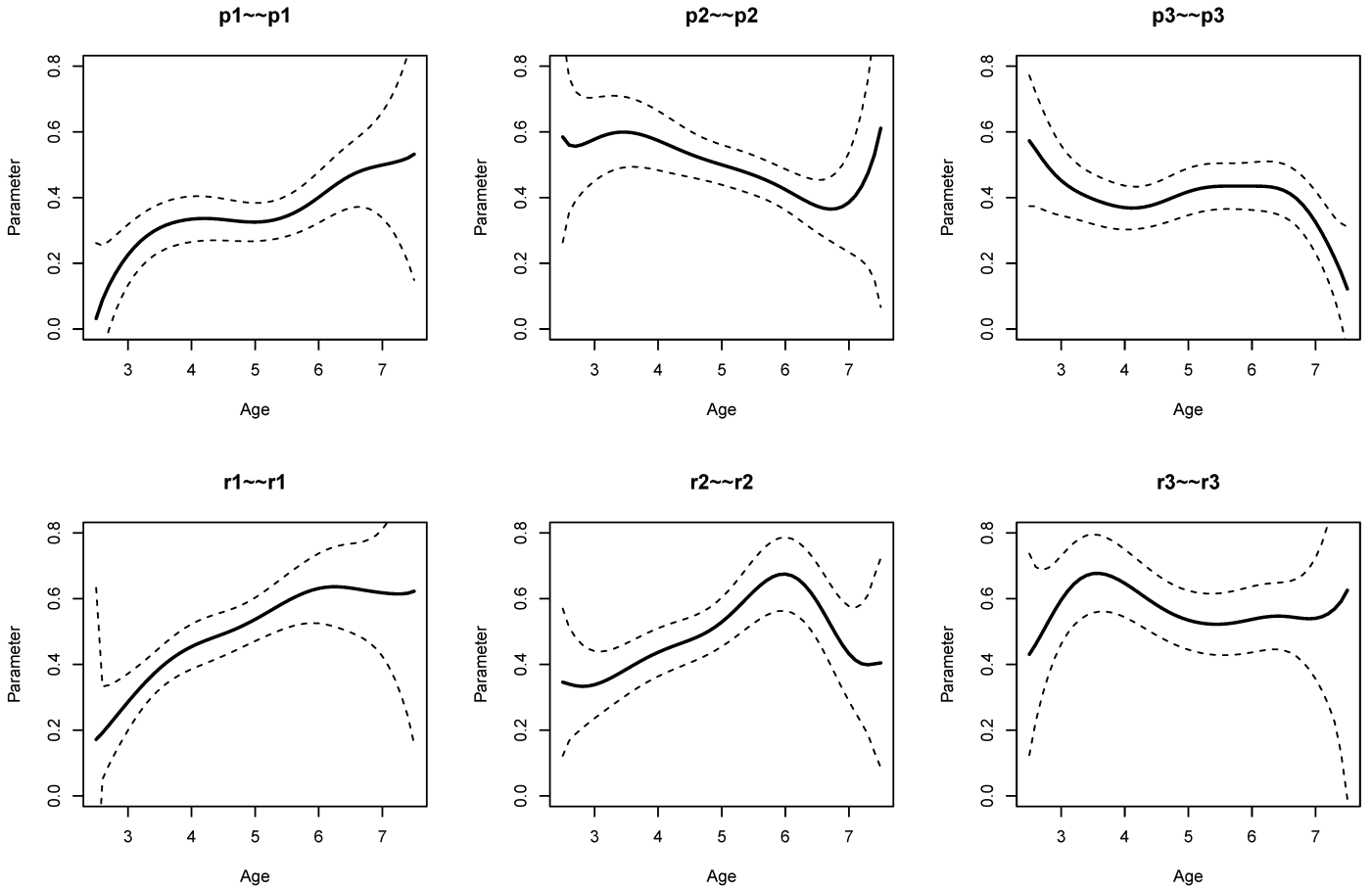

Figure 8. SON-R example: Parameter curves for residual variances.

Note that Hülür et al. (2011) used a pointwise LSEM approach instead a joint LSEM estimation approach. The identification of parameters in the covariance structure of factors was achieved in Hülür et al. (2011) by the constraint that the pointwise average of factor loadings equaled 1. Due to the different estimation approaches, it is expected that there are slight differences between our joint LSEM estimation approach and the original analysis in Hülür et al. (2011). The parameter curve of the correlation between the performance and the reasoning factors was similar in both analyses, with the exception that the factor correlation for small age values was much lower in the joint estimation approach, as displayed in Figure 6.

An anonymous reviewer wondered whether the factor correlation could be meaningfully interpreted if factor loadings did not show invariance across the moderator values. We argued elsewhere that measurement invariance would be a helpful but not a necessary condition for a meaningful interpretation of a factor correlation or a factor variance (see Robitzsch and Lüdtke 2023). In fact, a violation of measurement invariance only implies that results would change if a subset of indicators was used in the factor model. Because the SON-R instrument is held fixed in test administration and statistical analysis, this property of item selection invariance is not required. Of course, any identification constraint on factor loadings must be imposed to identify a factor correlation. The choice of identification constraint is somehow arbitrary. It could be invariance of all factor loadings, invariance of loadings of a subset of indicators, or a pointwise constraint of the average loadings (i.e., the average loading should be 1 for all indicators of a factor).

9. Discussion

In this article, we discussed the implementation of LSEM in the R package sirt. Joint LSEM estimation and two different significance tests for a test of parameter variation were introduced and evaluated through two simulation studies.

Simulation Study 1 demonstrated that the joint LSEM estimation method can be successfully applied to structural equation models whose parameters vary across different values of the moderator variable. It turned out that the bandwidth factor $h = 2$ can generally be recommended as a default choice. Notably, LSEM model parameters can be quite variable for small ($N = 250$) or moderate sample sizes ($N = 500$). In Simulation Study 2, two significance testing approaches for constant parameter curves were investigated: a test statistic based on the standard deviation of a parameter curve and a Wald-type test statistic. Both testing approaches rely on bootstrap samples for statistical inference. The standard-deviation-based test statistics had a higher power than the Wald test-type test statistic, but also came with an inflated type-I error rate. It is recommended to use the significance test based on the standard deviation with a significance level of 1% if a nominal significance level of 5% is required.

The application of LSEM in applied research can be regarded more as an exploratory than a confirmatory statistical method (Jacobucci 2022). Functional forms of parameter curves obtained with LSEM can be validated in other samples or future studies with more confirmatory approaches, such as moderated nonlinear factor analysis. We would like to emphasize that sufficiently large sample sizes are required in LSEM in order to allow a reliable interpretation of the obtained nonlinear parameter curves. Moreover, the true variability in parameter curves must be sufficiently large to have enough power to statistically detect the significant parameter variability. A statistical significance test on parameter curve regression coefficients in a moderated nonlinear factor analysis might have more power than a test based on the nonparametric LSEM method. Finally, moderated nonlinear factor analysis, if estimated by maximum likelihood, allows likelihood ratio tests for testing among nested models or using information criteria for model comparisons.

In this article, the moderator variable was exclusively age and a bounded variable. There might be applications in which the moderator differs from age, such as unbounded self-concept factor variables or ability values obtained from item response models (Basarkod et al. 2023). Because the metric of such variables is often arbitrary, it is advised to transform such moderators into a bounded metric. For example, the percentage ranks of an unbounded moderator variable could be utilized to obtain a bounded moderator variable.

If the moderator variable is an error-prone variable such as a factor variable or a scale score, an expected a posteriori (EAP) factor score estimate can be used as a moderator to obtain unbiased estimates of LSEM model parameters (Bartholomew et al. 2011).

As explained in Section 4, datasets with missing values should either be handled with pairwise deletion methods for computing sufficient statistics (i.e., the conditional covariance matrices) in LSEM or should be multiply imputed. The imputation model should be flexibly specified to represent the complex associations modeled with LSEM. For example, the moderator variable could be discretized into 5 or 10 distinct groups, and the resulting datasets should be separately imputed in the separate subdatasets. Statistical inference should be carried out that involves the multiply imputed datasets (Little and Rubin 2002).

Finally, we only discussed LSEM in the case of one moderator variable. With more than one moderator variable (Hartung et al. 2018), moderated nonlinear factor analysis might be easier to estimate because multivariate kernel functions for LSEM are difficult to estimate with sparse data.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Datasets and R code is available as supplementary material at https://osf.io/puaz9/?view_only=63ffb2fd30f5400e89c59d03366bf793 (accessed on 3 June 2023).

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CFA	confirmatory factor analysis
DGM	data-generating model
DIF	differential item functioning
LSEM	local structural equation model
ML	maximum likelihood
MNFA	moderated nonlinear factor analysis
SEM	structural equation model

References

Allemand, Mathias, Gabriel Olaru, and Patrick L. Hill. 2021. Age-related psychometrics and differences in gratitude and future time perspective across adulthood. *Personality and Individual Differences* 182: 111086. [CrossRef]

Allemand, Mathias, Gabriel Olaru, and Patrick L. Hill. 2022. Gratitude and future time perspective during the COVID-19 pandemic: Effects of age and virus worry. *The Journal of Positive Psychology* 17: 819–31. [CrossRef]

Arnold, Manuel, Andreas M. Brandmaier, and Manuel C. Voelkle. 2021. Predicting differences in model parameters with individual parameter contribution regression using the R package ipcr. *Psych* 3: 360–85. [CrossRef]

Arnold, Manuel, Daniel L. Oberski, Andreas M. Brandmaier, and Manuel C. Voelkle. 2020. Identifying heterogeneity in dynamic panel models with individual parameter contribution regression. *Structural Equation Modeling: A Multidisciplinary Journal* 27: 613–28. [CrossRef]

Bartholomew, David J., Martin Knott, and Irini Moustaki. 2011. *Latent Variable Models and Factor Analysis: A Unified Approach*. New York: Wiley. [CrossRef]

Basarkod, Geetanjali, Herbert W. Marsh, Baljinder K. Sahdra, Philip D. Parker, Jiesi Guo, Theresa Dicke, and Oliver Lüdtke. 2023. The dimensionality of reading self-concept: Examining its stability using local structural equation models. *Assessment* 30: 873–90. [CrossRef] [PubMed]

Bates, Timothy C., Hermine Maes, and Michael C. Neale. 2019. umx: Twin and path-based structural equation modeling in R. *Twin Research and Human Genetics* 22: 27–41. [CrossRef]

Bauer, Daniel J. 2017. A more general model for testing measurement invariance and differential item functioning. *Psychological Methods* 22: 507–26. [CrossRef]

Bentler, Peter M., and Ke-Hai Yuan. 2011. Positive definiteness via off-diagonal scaling of a symmetric indefinite matrix. *Psychometrika* 76: 119–23. [CrossRef]

Boker, Steven, Michael Neale, Hermine Maes, Michael Wilde, Michael Spiegel, Timothy Brick, Jeffrey Spies, Ryne Estabrook, Sarah Kenny, Timothy Bates, and et al. 2011. OpenMx: An open source extended structural equation modeling framework. *Psychometrika* 76: 306–17. [CrossRef]

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York: Wiley. [CrossRef]

Bollen, Kenneth A., and Walter R. Davis. 2009. Two rules of identification for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal* 16: 523–36. [CrossRef]

Bolsinova, Maria, and Dylan Molenaar. 2018. Modeling nonlinear conditional dependence between response time and accuracy. *Frontiers in Psychology* 9: 1525. [CrossRef]

Bolsinova, Maria, and Dylan Molenaar. 2019. Nonlinear indicator-level moderation in latent variable models. *Multivariate Behavioral Research* 54: 62–84. [CrossRef] [PubMed]

Boos, Dennis D., and Leonard A. Stefanski. 2013. *Essential Statistical Inference*. New York: Springer. [CrossRef]

Brandt, Naemi D., Johanna Drewelies, Sherry L. Willis, K. Warner Schaie, Nilam Ram, Denis Gerstorf, and Jenny Wagner. 2022. Beyond big five trait domains: Stability and change in personality facets across midlife and old age. *Journal of Personality*, *Epub ahead of print*. [CrossRef] [PubMed]

Bratt, Christopher, Dominic Abrams, Hannah J. Swift, Christin-Melanie Vauclair, and Sibila Marques. 2018. Perceived age discrimination across age in Europe: From an ageing society to a society for all ages. *Developmental Psychology* 54: 167–80. [CrossRef] [PubMed]

Breit, Moritz, Julian Preuß, Vsevolod Scherrer, Tobias Moors, and Franzis Preckel. 2023. Relationship between creativity and intelligence: A multimethod investigation of alternative theoretical assumptions in two samples of secondary school students. *Gifted Child Quarterly* 67: 95–109. [CrossRef]

Breit, Moritz, Martin Brunner, and Franzis Preckel. 2020. General intelligence and specific cognitive abilities in adolescence: Tests of age differentiation, ability differentiation, and their interaction in two large samples. *Developmental Psychology* 56: 364–84. [CrossRef]

Breit, Moritz, Martin Brunner, and Franzis Preckel. 2021. Age and ability differentiation in children: A review and empirical investigation. *Developmental Psychology* 57: 325–46. [CrossRef]

Breit, Moritz, Martin Brunner, Dylan Molenaar, and Franzis Preckel. 2022. Differentiation hypotheses of intelligence: A systematic review of the empirical evidence and an agenda for future research. *Psychological Bulletin* 148: 518–54. [CrossRef]

Briley, Daniel A., K. Paige Harden, and Elliot M. Tucker-Drob. 2015b. Genotype×cohort interaction on completed fertility and age at first birth. *Behavior Genetics* 45: 71–83. [CrossRef]

Briley, Daniel A., K. Paige Harden, Timothy C. Bates, and Elliot M. Tucker-Drob. 2015a. Nonparametric estimates of gene× environment interaction using local structural equation modeling. *Behavior Genetics* 45: 581–96. [CrossRef]

Browne, Michael W. 1974. Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal* 8: 1–24. Available online: https://bit.ly/3yviejm (accessed on 3 June 2023). [CrossRef]

Browne, Michael W., and Gerhard Arminger. 1995. Specification and estimation of mean-and covariance-structure models. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Edited by Gerhard Arminger, Clifford C. Clogg and Michael E. Sobel. Boston: Springer, pp. 185–249. [CrossRef]

Cox, Simon R., Stuart J. Ritchie, Elliot M. Tucker-Drob, David C. Liewald, Saskia P. Hagenaars, Gail Davies, Joanna M. Wardlaw, Catharine R. Gale, Mark E. Bastin, and Ian J. Deary. 2016. Ageing and brain white matter structure in 3513 UK Biobank participants. *Nature Communications* 7: 1–13. [CrossRef]

Curran, Patrick J., James S. McGinley, Daniel J. Bauer, Andrea M. Hussong, Alison Burns, Laurie Chassin, Kenneth Sher, and Robert Zucker. 2014. A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. *Multivariate Behavioral Research* 49: 214–31. [CrossRef] [PubMed]

de Mooij, Susanne M. M., Richard N. A. Henson, Lourens J. Waldorp, and Rogier A. Kievit. 2018. Age differentiation within gray matter, white matter, and between memory and white matter in an adult life span cohort. *Journal of Neuroscience* 38: 5826–36. [CrossRef]

Dong, Yixiao, and Denis Dumas. 2020. Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. *Personality and Individual Differences* 160: 109956. [CrossRef]

Dutton, Edward, and Emil Kirkegaard. 2022. The negative religiousness-IQ nexus is a Jensen effect on individual-level data: A refutation of Dutton et al.'s "The myth of the stupid believer". *Journal of Religion and Health* 61: 3253–75. [CrossRef] [PubMed]

Efron, Bradley, and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Boca Raton: CRC Press. [CrossRef]

Entringer, Theresa M., and Samuel D. Gosling. 2022. Loneliness during a nationwide lockdown and the moderating effect of extroversion. *Social Psychological and Personality Science* 13: 769–80. [CrossRef]

Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*. Thousand Oaks: Sage. Available online: https://bit.ly/38XUSX1 (accessed on 3 June 2023).

Gana, Kamel, Nedjem Eddine Boudouda, Thomas Salanova, and Guillaume Broc. 2023. What do EURO-D scores capture? Disentangling trait and state variances in depression symptoms across the adult life span in nine European nations. *Assessment*, *Epub ahead of print*. [CrossRef]

Gnambs, Timo. 2013. The elusive general factor of personality: The acquaintance effect. *European Journal of Personality* 27: 507–20. [CrossRef]

Gnambs, Timo, and Ulrich Schroeders. 2020. Cognitive abilities explain wording effects in the Rosenberg self-esteem scale. *Assessment* 27: 404–18. [CrossRef]

Gourieroux, Christian, Alain Monfort, and Alain Trognon. 1984. Pseudo maximum likelihood methods: Theory. *Econometrica* 52: 681–700. [CrossRef]

Grund, Simon, Oliver Lüdtke, and Alexander Robitzsch. 2022. Using synthetic data to improve the reproducibility of statistical results in psychological research. *Psychological Methods*, *Epub ahead of print*. [CrossRef] [PubMed]

Han, Kyunghee, Stephen M. Colarelli, and Nathan C. Weed. 2019. Methodological and statistical advances in the consideration of cultural diversity in assessment: A critical review of group classification and measurement invariance testing. *Psychological Assessment* 31: 1481–96. [CrossRef]

Hartung, Johanna, Laura E. Engelhardt, Megan L. Thibodeaux, K. Paige Harden, and Elliot M. Tucker-Drob. 2020. Developmental transformations in the structure of executive functions. *Journal of Experimental Child Psychology* 189: 104681. [CrossRef] [PubMed]

Hartung, Johanna, Martina Bader, Morten Moshagen, and Oliver Wilhelm. 2022. Age and gender differences in socially aversive ("dark") personality traits. *European Journal of Personality* 36: 3–23. [CrossRef]

Hartung, Johanna, Philipp Doebler, Ulrich Schroeders, and Oliver Wilhelm. 2018. Dedifferentiation and differentiation of intelligence in adults across age and years of education. *Intelligence* 69: 37–49. [CrossRef]

Hartung, Johanna, Sandy S. Spormann, Morten Moshagen, and Oliver Wilhelm. 2021. Structural differences in life satisfaction in a US adult sample across age. *Journal of Personality* 89: 1232–51. [CrossRef]

Hildebrandt, Andrea, Oliver Lüdtke, Alexander Robitzsch, Christopher Sommer, and Oliver Wilhelm. 2016. Exploring factor model parameters across continuous variables with local structural equation models. *Multivariate Behavioral Research* 51: 257–78. [CrossRef]

Hildebrandt, Andrea, Oliver Wilhelm, and Alexander Robitzsch. 2009. Complementary and competing factor analytic approaches for the investigation of measurement invariance. *Review of Psychology* 16: 87–102.

Hildebrandt, Andrea, Oliver Wilhelm, Grit Herzmann, and Werner Sommer. 2013. Face and object cognition across adult age. *Psychology and Aging* 28: 243–48. [CrossRef] [PubMed]

Hildebrandt, Andrea, Werner Sommer, Grit Herzmann, and Oliver Wilhelm. 2010. Structural invariance and age-related performance differences in face cognition. *Psychology and Aging* 25: 794–810. [CrossRef]

Holland, Paul W., and Howard Wainer, eds. 1993. *Differential Item Functioning: Theory and Practice*. Hillsdale: Lawrence Erlbaum. [CrossRef]

Hülür, Gizem, Oliver Wilhelm, and Alexander Robitzsch. 2011. Intelligence differentiation in early childhood. *Journal of Individual Differences* 32: 170–79. [CrossRef]

Huth, Karoline B. S., Lourens J. Waldorp, Judy Luigjes, Anneke E. Goudriaan, Ruth J. van Holst, and Marten Marsman. 2022. A note on the structural change test in highly parameterized psychometric models. *Psychometrika* 87: 1064–80. [CrossRef]

Jacobucci, Ross. 2022. A critique of using the labels confirmatory and exploratory in modern psychological research. *Frontiers in Psychology* 13: 1020770. [CrossRef] [PubMed]

Jiang, Bei, Adrian E. Raftery, Russell J. Steele, and Naisyin Wang. 2021. Balancing inferential integrity and disclosure risk via model targeted masking and multiple imputation. *Journal of the American Statistical Association* 117: 52–66. [CrossRef]

Jokić-Begić, Nataša, Una Mikac, Doris Čuržik, and Claire Sangster Jokić. 2019. The development and validation of the short cyberchondria scale (SCS). *Journal of Psychopathology and Behavioral Assessment* 41: 662–76. [CrossRef]

Jöreskog, Karl G., Ulf H. Olsson, and Fan Y. Wallentin. 2016. *Multivariate Analysis with LISREL*. Basel: Springer. [CrossRef]

Kaltwasser, Laura, Una Mikac, Vesna Buško, and Andrea Hildebrandt. 2017. No robust association between static markers of testosterone and facets of socio-economic decision making. *Frontiers in Behavioral Neuroscience* 11: 250. [CrossRef]

Klieme, Katrin E., and Florian Schmidt-Borcherding. 2023. Lacking measurement invariance in research self-efficacy: Bug or feature? *Frontiers in Education* 8: 1092714. [CrossRef]

Kolbe, Laura, Dylan Molenaar, Suzanne Jak, and Terrence D. Jorgensen. 2022. Assessing measurement invariance with moderated nonlinear factor analysis using the R package OpenMx. *Psychological Methods*, *Epub ahead of print*. [CrossRef]

Kolenikov, Stanislav. 2010. Resampling variance estimation for complex survey data. *The Stata Journal* 10: 165–99. [CrossRef]

Kolenikov, Stanislav. 2011. Biases of parameter estimates in misspecified structural equation models. *Sociological Methodology* 41: 119–57. [CrossRef]

Krasko, Julia, and Till Kaiser. 2023. Die Dunkle Triade in einer deutschen repräsentativen Stichprobe [The dark triad in a German representative sample]. *Diagnostica* 69: 1–13. [CrossRef]

Leitgöb, Heinz, Daniel Seddig, Tihomir Asparouhov, Dorothée Behr, Eldad Davidov, Kim De Roover, Suzanne Jak, Katharina Meitinger, Natalja Menold, Bengt Muthén, and et al. 2023. Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research* 110: 102805. [CrossRef] [PubMed]

Little, Roderick J. A., and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: Wiley. [CrossRef]

Liu, Xinyang, Mattis Geiger, Changsong Zhou, and Andrea Hildebrandt. 2022. Individual differences in white matter microstructure of the face processing brain network are more differentiated from global fibers with increasing ability. *Scientific Reports* 12: 14075. [CrossRef]

Lodi-Smith, Jennifer, Jonathan D. Rodgers, Valeria Marquez Luna, Sarah Khan, Caleb J. Long, Karl F. Kozlowski, James P. Donnelly, Christopher Lopata, and Marcus L. Thomeer. 2021. The relationship of age with the autism-spectrum quotient scale in a large sample of adults. *Autism in Adulthood* 3: 147–56. [CrossRef]

Madole, James W., Mijke Rhemtulla, Andrew D. Grotzinger, Elliot M. Tucker-Drob, and K. Paige Harden. 2019. Testing cold and hot cognitive control as moderators of a network of comorbid psychopathology symptoms in adolescence. *Clinical Psychological Science* 7: 701–18. [CrossRef] [PubMed]

Mellenbergh, Gideon J. 1989. Item bias and item response theory. *International Journal of Educational Research* 13: 127–43. [CrossRef]

Meredith, William. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58: 525–43. [CrossRef]

Merkle, Edgar C., and Achim Zeileis. 2013. Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika* 78: 59–82. [CrossRef] [PubMed]

Millsap, Roger E. 2011. *Statistical Approaches to Measurement Invariance*. New York: Routledge. [CrossRef]

Molenaar, Dylan. 2021. A flexible moderated factor analysis approach to test for measurement invariance across a continuous variable. *Psychological Methods* 26: 660–79. [CrossRef]

Molenaar, Dylan, and Conor V. Dolan. 2012. Substantively motivated extensions of the traditional latent trait model. *Netherlands Journal of Psychology* 67: 48–57. Available online: https://bit.ly/3Tu30oV (accessed on 3 June 2023).

Molenaar, Dylan, Conor V. Dolan, and Han L. J. van der Maas. 2011. Modeling ability differentiation in the second-order factor model. *Structural Equation Modeling: A Multidisciplinary Journal* 18: 578–94. [CrossRef]

Molenaar, Dylan, Conor V. Dolan, and Norman D. Verhelst. 2010a. Testing and modelling non-normality within the one-factor model. *British Journal of Mathematical and Statistical Psychology* 63: 293–317. [CrossRef] [PubMed]

Molenaar, Dylan, Conor V. Dolan, Jelte M. Wicherts, and Han L. J. van der Maas. 2010b. Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence* 38: 611–24. [CrossRef]

Molenaar, Dylan, Natasa Kő, Sandor Rózsa, and Andrea Mészáros. 2017. Differentiation of cognitive abilities in the WAIS-IV at the item level. *Intelligence* 65: 48–59. [CrossRef]

Nowok, Beata, Gillian M. Raab, and Chris Dibben. 2016. synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software* 74: 1–26. [CrossRef]

Oberski, Daniel L. 2013. Individual differences in structural equation model parameters. *arXiv* arXiv:1304.3608.

Olaru, Gabriel, and Mathias Allemand. 2022. Correlated personality change across time and age. *European Journal of Personality* 36: 729–49. [CrossRef]

Olaru, Gabriel, and Kristin Jankowsky. 2022. The HEX-ACO-18: Developing an age-invariant HEXACO short scale using ant colony optimization. *Journal of Personality Assessment* 104: 435–46. [CrossRef]

Olaru, Gabriel, Alexander Robitzsch, Andrea Hildebrandt, and Ulrich Schroeders. 2020. Local structural equation modeling for longitudinal data. *PsyArXiv*. April 24. [CrossRef]

Olaru, Gabriel, Alexander Robitzsch, Andrea Hildebrandt, and Ulrich Schroeders. 2022. Examining moderators of vocabulary acquisition from kindergarten through elementary school using local structural equation modeling. *Learning and Individual Differences* 95: 102136. [CrossRef]

Olaru, Gabriel, Ulrich Schroeders, Johanna Hartung, and Oliver Wilhelm. 2019. Ant colony optimization and local weighted structural equation modeling. A tutorial on novel item and person sampling procedures for personality research. *European Journal of Personality* 33: 400–19. [CrossRef]

Olaru, Gabriel, Ulrich Schroeders, Oliver Wilhelm, and Fritz Ostendorf. 2019. 'Grandpa, do you like roller coasters?': Identifying age-appropriate personality indicators. *European Journal of Personality* 33: 264–78. [CrossRef]

Panayiotou, Margarita, Johanna C. Badcock, Michelle H. Lim, Michael J. Banissy, and Pamela Qualter. 2022. Measuring loneliness in different age groups: The measurement invariance of the UCLA loneliness scale. *Assessment*, *Epub ahead of print*. [CrossRef]

Putnick, Diane L., and Marc H Bornstein. 2016. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review* 41: 71–90. [CrossRef]

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team. Available online: https://www.R-project.org/ (accessed on 15 March 2023).

Reiter, Jerome P. 2023. Synthetic data: A look back and a look forward. *Transactions on Data Privacy* 16: 15–24. Available online: http://www.tdp.cat/issues21/tdp.a457a22.pdf (accessed on 3 June 2023).

Robitzsch, Alexander. 2023a. Model-robust estimation of multiple-group structural equation models. *Algorithms* 16: 210. [CrossRef]

Robitzsch, Alexander. 2023b. Sirt: Supplementary Item Response theory Models. R Package Version 3.13-228. Available online: https://CRAN.R-project.org/package=sirt (accessed on 11 August 2023) .

Robitzsch, Alexander, and Oliver Lüdtke. 2023. Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *Epub ahead of print*. [CrossRef]

Rosseel, Yves. 2012. lavaan: An R package for structural equation modeling. *Journal of Statistical Software* 48: 1–36. [CrossRef]

Schroeders, Ulrich, and Malte Jansen. 2022. Science self-concept–more than the sum of its parts? *The Journal of Experimental Education* 90: 435–51. [CrossRef]

Schroeders, Ulrich, Stefan Schipolowski, and Oliver Wilhelm. 2015. Age-related changes in the mean and covariance structure of fluid and crystallized intelligence in childhood and adolescence. *Intelligence* 48: 15–29. [CrossRef]

Seifert, Ingo S., Julia M. Rohrer, Boris Egloff, and Stefan C. Schmukle. 2022. The development of the rank-order stability of the big five across the life span. *Journal of Personality and Social Psychology* 122: 920–41. [CrossRef] [PubMed]

Shapiro, Alexander. 2012. Statistical inference of covariance structures. In *Current Topics in the Theory and Application of Latent Variable Models*. Edited by M. C. Edwards and R. C. MacCallum. London: Routledge, pp. 222–40. [CrossRef]

Silverman, Bernard W. 1986. *Density Estimation for Statistics and Data Analysis*. Boca Raton: CRC Press. [CrossRef]

Stefanski, Leonard A., and Dennis D. Boos. 2002. The calculus of M-estimation. *The American Statistician* 56: 29–38. [CrossRef]

Tucker-Drob, Elliot M. 2009. Differentiation of cognitive abilities across the life span. *Developmental Psychology* 45: 1097–118. [CrossRef] [PubMed]

Van den Akker, Alithe L., Daniel A. Briley, Andrew D. Grotzinger, Jennifer L. Tackett, Elliot M. Tucker-Drob, and K. Paige Harden. 2021. Adolescent big five personality and pubertal development: Pubertal hormone concentrations and self-reported pubertal status. *Developmental Psychology* 57: 60–72. [CrossRef]

Wagner, Jenny, Oliver Lüdtke, and Alexander Robitzsch. 2019. Does personality become more stable with age? Disentangling state and trait effects for the big five across the life span using local structural equation modeling. *Journal of Personality and Social Psychology* 116: 666–80. [CrossRef]

Wang, Ting, Edgar C. Merkle, and Achim Zeileis. 2014. Score-based tests of measurement invariance: Use in practice. *Frontiers in Psychology* 5: 438. [CrossRef] [PubMed]

Watrin, Luc, Ulrich Schroeders, and Oliver Wilhelm. 2022. Structural invariance of declarative knowledge across the adult lifespan. *Psychology and Aging* 37: 283–97. [CrossRef]

Weiss, Selina, Diana Steger, Ulrich Schroeders, and Oliver Wilhelm. 2020. A reappraisal of the threshold hypothesis of creativity and intelligence. *Journal of Intelligence* 8: 38. [CrossRef]

White, Halbert. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50: 1–25. [CrossRef]

Whitley, Elise, Ian J. Deary, Stuart J. Ritchie, G. David Batty, Meena Kumari, and Michaela Benzeval. 2016. Variations in cognitive abilities across the life course: Cross-sectional evidence from understanding society: The UK household longitudinal study. *Intelligence* 59: 39–50. [CrossRef] [PubMed]

Yuan, Ke-Hai, and Peter M. Bentler. 2007. Structural equation modeling. In *Handbook of Statistics, Vol. 26: Psychometrics*. Edited by C. Randhakrishna Rao and Sandip Sinharay. pp. 297–358. [CrossRef]

Zheng, Anqing, Daniel A. Briley, Margherita Malanchini, Jennifer L. Tackett, K. Paige Harden, and Elliot M. Tucker–Drob. 2019. Genetic and environmental influences on achievement goal orientations shift with age. *European Journal of Personality* 33: 317–36. [CrossRef] [PubMed]

Journal of
Intelligence

MDPI

Article

Flexibility to Change the Solution: An Indicator of Problem Solving That Predicted 9th Grade Students' Academic Achievement during Distance Learning, in Parallel to Reasoning Abilities and Parental Education

Liena Hacatrjana

Department of Psychology, Faculty of Education, Psychology and Art, University of Latvia, Riga, Imantas 7. linija-1, LV-1083 Riga, Latvia; liena.hacatrjana@lu.lv

Abstract: The relation between academic achievement and various measurements of cognitive abilities, problem-solving skills and self-managed learning has been established in the research before the COVID-19 pandemic and distance learning. The aim of the current research was to analyze the extent to which these aspects predicted the educational achievement of 9th grade students (mean age 15.4 years) during distance learning, when students had to do relatively more tasks independently, organize their daily learning and deal with problems on their own. Relations between self-assessed problem-solving skills, self-management skills, tests of reasoning abilities and the results of diagnostic tests in Mathematics and Latvian were analyzed for n = 256 and n = 244 students, respectively. The results show that: (1) diagnostic test results in Mathematics are best predicted by the parental education level, fluid nonverbal reasoning and verbal reasoning; (2) the best predictors for the results in the diagnostic test in Latvian are parental education, flexibility to change the solution, fluid nonverbal reasoning and verbal reasoning; (3) self-management cannot significantly predict the results of either of the two tests, although it correlates to the results of the tests in both Mathematics and Latvian; (4) only one of the aspects of problem-solving, flexibility to change the solution, can significantly predict results in diagnostic tests. The results confirm the significance of cognitive abilities as an important predictor of academic achievement, as well as the role of parents' education level. The results also suggest that the flexibility to change the solution, an aspect of problem-solving, might play a role in students' success in academic tests.

Keywords: academic achievement; COVID-19; distance learning; cognitive abilities; self-assessed skills; problem-solving; self-management skills; parental education

Citation: Hacatrjana, Liena. 2022. Flexibility to Change the Solution: An Indicator of Problem Solving That Predicted 9th Grade Students' Academic Achievement during Distance Learning, in Parallel to Reasoning Abilities and Parental Education. *Journal of Intelligence* 10: 7. https://doi.org/10.3390/jintelligence10010007

Received: 9 November 2021
Accepted: 24 January 2022
Published: 27 January 2022

1. Introduction

During the COVID-19 pandemic, more than 1.6 billion children worldwide faced disruptions in face-to-face education, and many schools continued the educational process with distance learning (OECD 2020). To many students it was a new situation and their first experience with distance learning, bringing new challenges that could be considered as problems that needed to be solved daily. Different approaches to distance learning were applied in various countries around the world (Reimers and Schleicher 2020). In Latvia, a country among those with a high number of days of remote learning, mixed forms of learning were implemented (both synchronous and asynchronous) with online video lessons and with assigned tasks to be done individually at home (Ministry of Education and Science of Latvia 2020) indicating that the ability to work independently was demanded from students. Teachers were forced to swiftly adapt to using new technologies and using pedagogical techniques that worked online, but not all teachers were immediately ready for this: 25% of teachers reported that they had not organized any online lesson in the spring of 2020, when the first remote learning period was implemented, indicating that they had

sent materials to students that had to be learned independently. Students were forced to learn on their own via online lessons with teachers or without direct online communication with teachers. During the first wave of pandemic, in spring 2020, about half of the students in Latvia reported that they lacked teachers' explanations and motivation while at home and not in their classroom, and they felt stressed and unsure whether they would finish all tasks in time (Ministry of Education and Science of Latvia 2020), which indicates that an extra effort was asked of them.

The pandemic period and the distance learning have raised questions globally about which skills, abilities and other factors (e.g., external factors such as support from parents or teachers) are crucial for students to maintain their academic performance and well-being as much as possible during this time (Rosen et al. 2021; Hacatrjana 2021a), and some recent data have indicated a decrease in students' academic performance that is probably due to the pandemic (Engzell et al. 2021). In this study, the focus is particularly on the individual aspects related to the students' own skills and abilities to deal with the new situation. Students (most studies were performed in primary schools or in high schools) have reported in questionnaires that their ability to plan their time has helped them during distance learning, while also indicating feeling stress regarding the management of assignments on time and having insufficient planning skills, self-organization and management skills, which caused difficulties with remote learning (Scott et al. 2021; Ministry of Education and Science of Latvia 2020; Rogers et al. 2021; Hacatrjana 2021a). This means that during the pandemic and distance learning many students were aware that they lacked some skills to learn effectively. In addition, the pandemic situation and the unprecedented distance learning were essentially a new challenge for most students, and thus they needed to adapt to the new situation and use their problem-solving skills to cope with it and to study independently in a successful manner. In Latvia, a change in the curriculum was recently introduced in schools, and problem-solving skills and self-regulated learning skills are among the transversal skills that are deemed important for students in Latvia and that should be further developed at schools (Cabinet of Ministers Republic of Latvia 2018). In addition, previous PISA results have indicated that the results on problem-solving skills in this country are below the OECD average (OECD 2017). During the pandemic a study with high-school students in Latvia showed that students with higher self-reported problem-solving skills were less stressed about distance learning (Hacatrjana 2021a). All these previously obtained results suggest that problem-solving skills and self-management skills are essential for students to adapt to new circumstances and to maintain their academic achievement during distance learning in the COVID-19 pandemic.

The close link between various measurements of academic achievement and indicators of cognitive abilities has been well established in the research literature (way before the pandemic), proving that cognitive abilities predict academic achievement to a great extent (e.g., Frey 2019; Kampa et al. 2021). Indicators of other skills show a relation to academic achievement as well, both with GPA and SAT tests. For example, problem-solving tests predict academic achievement (Greiff et al. 2014; Greiff et al. 2013). Self-management indicators, mostly assessed with self-assessment type questionnaires, are also related to academic achievement (Pintrich et al. 1993; Zimmerman and Martinez-Pons 1988; Abd-El-Fattah 2010; Veenman et al. 2014). Problem-solving skills (assessed via various methodological approaches) and cognitive ability also show interrelationships (e.g., Chuderski and Jastrzębski 2018; Kretzschmar et al. 2017; Ellis et al. 2021). In general, these results indicate that there is a set of cognitive abilities and additional skills which together can predict a student's learning performance in regular learning settings. Given the differences from the usual learning environment and the format in which most research has taken place in this field, it is important to explore the extent to which cognitive abilities and additional skills—problem solving and self-management skills—predict student performance during distance learning in the pandemic. The aim of the current research is to examine whether problem-solving skills and self-management skills, in parallel with tested cognitive abilities, can predict the results of 9th grade students' diagnostic tests (an indicator of academic

achievement) during the distance learning period due to the COVID-19 pandemic, as it is discussed that both these skills are important when studying independently.

1.1. Problem-Solving Skills and Self-Management Skills: Important for Studying Independently

Assuming that problem-solving skills and self-management skills are important for students during distance learning (Hacatrjana 2021a), it is useful to unravel them in more detail in the context of this study. Regarding the research of problem-solving skills, there are several approaches in Psychology that differ based on their theoretical framework and methodology (e.g., Frensch and Funke 1995; Heppner and Petersen 1982; OECD 2013, and others). In addition, they are also defined as important skills in the education field in many countries, which seek to teach them to students (e.g., in Latvia, Cabinet of Ministers Republic of Latvia 2018). Most researchers in Psychology state that problem-solving consists of several underlying processes, often similar to the original ideas of George Polya proposed many years ago: (1) understanding the problem, (2) devising a plan, (3) carrying out the plan and (4) looking back (Polya 1957). One of the modern approaches that focuses on studying the abilities of complex problem solving empirically defines that there are two main processes underlying problem-solving: (1) knowledge acquisition and (2) knowledge application (Fischer et al. 2012). In the global PISA educational assessment problem-solving is considered to consist of several processes: (1) exploring and understanding; (2) representing and formulating; (3) planning and executing; (4) monitoring and reflecting; and reasoning is used during the process of problem-solving (OECD 2013). Other approaches focus on the self-assessment of the attitudes and experience in problem-solving (for example, Heppner and Petersen 1982), and problem-solving processes in specific fields—for example, in Mathematics (Verschaffel and Corte 1993). What the different approaches have in common is that the aspects of problem-solving skills are applied when facing a situation or a task which cannot be solved by an automated action, and often a clear and good solution is not immediately known, especially when facing new problems and situations.

Problem-solving skills in the context of the current research are defined by the author as a set of skills, habits and operations that help individuals (e.g., students), when facing a new task or problem, to successfully explore and understand the key concepts involved in the problem, to be able to come up with possible solutions, to implement a solution, to be able to realize if the solution is not appropriate and react accordingly (thus being flexible in the process of solving the problem), and to evaluate the result and process of problem solving. Problem-solving skills are here operationalized by the self-assessment of two aspects of problem solving: (1) Solution development and evaluation and (2) Flexibility to change the solution (see Methods section), indicating that during problem solving it is important to come up with possible solutions and evaluate the result afterwards, as well as being flexible to change the chosen solution strategy if it is not suitable. Flexibility is a variable also studied in the field of mathematical problem solving and it is related to academic achievement (e.g., Hästö et al. 2019). It includes the knowledge of various possible strategies and the ability to implement the appropriate option. These ideas from the field of mathematics could be transferred to problem solving in general, meaning that flexibility in problem solving indicates the ability to choose between the options a person can think of and to apply the most appropriate solution during the process of problem solving.

Self-management is a process that is involved in self-regulated learning (Zimmerman 2008), an important concept in modern educational approaches, and especially crucial during distance learning. Self-regulated learning is a very broad concept that consists of several important aspects. Both metacognitive processes and the ability to organize oneself practically are important for a person to become good at self-regulated learning. During the learning process students use their metacognition to proactively think, perform and self-reflect (Carter et al. 2020), which is assumed to lead to good self-management. According to the model by Garrisson (1997), self-management and self-organization are an important part of the broader concept of self-regulated learning, and they relate to how

the activities associated with learning are carried out and controlled, such as how all the necessary resources are managed.

In the current research, self-management skills are defined by the author as a set of skills and habits necessary to (1) successfully organize one's tasks, time and resources and (2) be able to understand conceptually and clearly what has to be done in a certain period of time and why it has to be done (motivational aspect). Thus, self-management is here focused mainly on one's practical organization, during the process of learning, also keeping the motivational aspect and focusing on the goal, as in the original ideas by Garrisson (1997).

The rationale of the current study is that it is important to assess how problem-solving skills and self-management skills are related to academic outcomes during unprecedented events such as the COVID-19 pandemic to make conclusions and develop further hypotheses about these skills in education (as a part of the curriculum). If these skills are shown to be important predictors during the pandemic, then we can further assume that they are indeed important skills to develop at schools because they might help students to adapt to any other unprecedented events that may come in the future.

1.2. Methodological Considerations Regarding the Relation between Self-Assessed Skills and Test Results

The analysis of the relationship between self-management, problem-solving skills, cognitive abilities and academic achievement should consider the research methodology—whether the skills are assessed by tests or by self-assessment methods. Researchers have proven that computer-based problem-solving skill tests have shown high correlations with cognitive abilities for high-school students (e.g., Kretzschmar et al. 2017) and a high ability to predict academic performance in primary and high school (e.g., Greiff et al. 2013). Some studies show that self-assessment measurements for problem solving also tend to show statistically significant relationships with cognitive test scores in the high-school population (e.g., Nota et al. 2009), and that self-assessed self-directed learning skills are related to academic performance, measured in an undergraduate sample (Tekkol and Demirel 2018).

When assessing the suitability of the self-assessment methods in educational settings in general, it becomes clear that research results on self-assessment accuracy are not consistent (Brown and Harris 2013), with some previous research showing that students' self-assessments correlate with the grading by teachers in schools (Sanchez et al. 2017). However, in some research, gender and other differences are reported regarding the accuracy of self-assessment—for example, the tendency for undergraduate students with higher marks to rate themselves more precisely compared to students with lower marks (González-Betancor et al. 2019; Kim et al. 2010). In a study with 9th grade students from Latvia, it was found that self-efficacy in mathematics was higher for boys (Kvedere 2014). Self-estimates of intelligence and test scores are low to moderately correlated in research with various samples (Rammstedt and Rammsayer 2002; Furnham and Grover 2020). Research also shows that undergraduate students with lower performance are more likely to overestimate themselves but are aware of the possible inaccuracy, and students with higher performance are more accurate in their self-evaluation (Miller and Geraci 2011). In general, the importance of self-assessment as a means for development of skills is discussed in the literature, and there is a tendency to increasingly include self-evaluation in the process of learning (Andrade 2019; Vasileiadou and Karadimitriou 2021). Research results indicate that there might be flaws in the precision of the self-assessment of one's skills, especially if the self-assessment affects the final mark (Andrade 2019), which was not the case in the current study. The author of this research used several methods to minimize the possible flaws of self-assessment: first, the participation in the study was anonymous, and students were encouraged to answer truly to themselves, not thinking about any "right or wrong answers"; and second, the questionnaire used in the study included indicators of specific operations that characterize problem solving and self-management that were clearly defined, and students had to evaluate how often they performed such activities.

Thus, the ratings were based on the frequency of an action, not on the agreement with a statement.

1.3. Academic Achievement Assessments during Distance Learning in the COVID-19 Pandemic

Another important issue that must be elaborated regarding the topic of this article is the practice and challenges of measuring academic achievement during the pandemic and the distance learning that was implemented in most countries as a response to the spread of the SARS-CoV-2 virus (OECD 2020). During this complicated time, there were different approaches implemented regarding students' assessment (Kuhfeld et al. 2020; Thorn and Vincent-Lancrin 2021). For example, summed scores by teachers were implemented in Ireland (Doyle et al. 2021). In Latvia the traditional exams at the end of primary school after the 9th grade have always been important and might determine one's chance of getting into a high school. However, during the COVID-19 pandemic a decision was made by the government that the usual exams after the 9th grade would be replaced with "diagnostic assessment tests" in the same taught subjects (mandatory for Mathematics and Latvian, optional for English, Sciences and History) (Cabinet of Ministers Republic of Latvia 2021). The diagnostic tests would not affect students' final grading at the end of primary school and further opportunities to join a high school or another schooling option. These tests would be similar in their content and level of difficulty to exams, and all 9th graders in Latvia took the same tests at the same time period, thus making this score appropriate for directly comparing the results of students in different schools. The results of these diagnostic tests were used in the current study as an indicator of the academic achievement of students.

1.4. Focus of the Current Article

This paper aims to determine the best predictors of diagnostic test results and the extent to which problem-solving skills and self-management skills were able to predict the results of diagnostic tests during distance learning in the pandemic situation, given that they were (in this case) determined with self-assessment methods, along with cognitive abilities that were assessed with test tasks and parental education level. As discussed before, academic achievement is proved to be closely related to indicators of cognitive abilities (e.g., Frey 2019; Kampa et al. 2021), and therefore it is assumed that in the current study the indicators of cognitive abilities should also significantly predict academic achievement. The significant relation between parental education level and academic achievement is also established in previous research (e.g., Idris et al. 2020). Thus, it is assumed that parental education should be an important predictor for academic achievement also in the current study. A recent study confirms the importance of parental education to grades, but it is concluded that intelligence is a more important predictor than the whole socioeconomic status measurement (Flores-Mendoza et al. 2021). It is assumed that problem-solving and self-management skills could play an important role in how well students were able to maintain their academic performance even during distance learning (Hacatrjana 2021a). This means that, hypothetically, if a student has good grades and high problem-solving skills, then he or she should also be able to deal well with studying in a new, unprecedented problem situation—distance learning. It is similar with the self-management and self-organization skills—if the student is doing well at school and has these skills highly developed, then it is easier for him or her to cope with distance learning, and vice versa—if a student has generally good grades, but he or she lacks self-management skills or problem-solving skills, then the distance learning process could have a greater impact on a student's performance, and academic achievement may be lower due to the lack of the skills to organize oneself and deal with problems.

The main question is: if we assume that problem-solving skills and self-management skills are indeed important to successfully cope with distance learning, will it show in the results of students' academic outcomes during this period? The aim of the current study is to examine this assumption, taking into account that these skills were assessed

with self-assessment methods. The research questions posed in the current study are: (1) What are the best predictors of the results of students' diagnostic tests at the end of 9th grade during the distance learning in the COVID-19 pandemic? (2) To what extent do the self-assessed problem-solving skills and self-management skills predict the results of diagnostic school tests of the mentioned population?

2. Materials and Methods

2.1. Sample

The data of n = 652 students in the 9th grade from general education schools in Latvia (359 females, 293 males), aged 14 to 17 years (M = 15.41, SD = 0.53), were gathered; the sample size used in the regression analysis is smaller due to a smaller amount of some of the indicators obtained from schools; in such cases, the precise amount of analyzed cases is reported within the results.

2.2. Measurements

(1) Problem-solving skills were evaluated with a problem-solving questionnaire, a self-assessment method with 10 items comprising two scales, that were named: (1) Solution development and evaluation (6 items) and (2) Flexibility to change the solution (4 items), that originally showed an internal consistency of, respectively, α = 0.79 and α = 0.71 (Hacatrjana 2021b). Each item had to be rated on a scale from "Never" to "Always" (0 to 5 points) based on how often a student performed the mentioned activity (item examples: "When solving a situation or doing a task, I change my solution if I understand that it is not appropriate", "When I have finished a task, I think about what worked well and what didn't."). The scale "Flexibility to change the solution" is significantly correlated to the results of nonverbal and verbal reasoning tests (r = 0.22 and r = 0.25, $p < 0.01$, respectively), indicating its validity, but statistically significant correlations are not found with the scale "Solution development and evaluation". Both scales of the questionnaire are significantly correlated (r = 0.46, $p < 0.01$).

(2) Self-management skills were assessed with the self-management questionnaire that is used for the purpose of self-assessing students' skills to manage and organize themselves and their learning. It consists of six items (for example, "I write down all the tasks in a certain place", "If I lose motivation at some point, I remind myself why it was important for me to do it"), that originally showed an internal consistency of α = 0.77. Each item had to be rated on a scale from "Never" to "Always" (0 to 5 points) based on how often a student performed such an action (Hacatrjana 2021b). In the current study the Self-management scale is negatively correlated to the students' self-evaluations of their perceived difficulty to deal with distance learning (r = -0.12, $p < 0.01$), indicating the validity of the scale.

(3) Fluid nonverbal reasoning was measured with a short version (10 items) of the Sandia Matrices test (see Harris et al. 2020; Matzen et al. 2010), that assesses reasoning abilities with typical figural matrices tasks where one has to understand the patterns in a set of drawings and choose the most appropriate answer (a drawing that continues the pattern) from eight answer options. The internal consistency of the test, measured with Chronbach's alpha, was α = 0.72. Each answer is rated with 0 or 1 point.

(4) Verbal reasoning was assessed with a short version of the Verbal analogies test (10 items) that has been previously developed and used in the research with students (Kretzschmar et al. 2017). In the test, one pair of words and the first word of the second pair is given (for example, "snow—to ski" and "ice—..."), and the participant has to understand the type of relationship for these words and write an answer to the second pair of words. The internal consistency of the test, measured with Chronbach's alpha, was α = 0.81. Each answer is rated with 0 or 1 point.

(5) Academic achievement was measured by gathering several indicators from schools: results in diagnostic tests at the end of the 9th grade in Mathematics, Latvian and

English. The tests were taken by students online during the pandemic, and each test was administered on a specific date set by the state. The test was exactly the same for all students in the country. It must be noted that not all students took all of the tests (some are optional, e.g., English), and not all schools provided the researcher with the necessary anonymized data; thus, the amount of available data is smaller for these test results compared to the data from other measurements. The exact amount of data analyzed is shown further in the results section. In each test a student can get from zero to a maximum of 100 points.

(6) Additional questions on experience and attitudes during distance learning were asked to students: for example, to rate their perceived difficulty to deal with the distance learning situation, to assess whether the technological means available to them were sufficient for studying. Students had to rate these questions on a Likert scale with 0 to 5 points. It was also asked if a student had been to an individual consultation with a teacher (individual face-to-face consultations were allowed as an exception at that period of time for students facing difficulties).

(7) Demographic questions were asked: gender, age, the level of parental education (from "1-Finished Primary school" to "6-Doctoral degree"). Each student wrote the individual code that was assigned by the school for each student to ensure confidentiality.

2.3. Procedure

The data collection was carried out in close collaboration with each participating school, in two rounds: (1) Students filled out the tests and questionnaires online. Students from each class joined a specifically scheduled online lesson on a platform typically used by the particular school in the period of distance learning (platforms "Microsoft Teams" and "Zoom" were most commonly used). Students were first informed about the study and instructed, and then they went to the testing site, on the internet, where they completed surveys and tests in 40–50 min. A link to the tests was given to the students at the beginning of the testing. The instructor remained connected to the online lesson to answer technical questions, if any came up. Pupils were asked to talk and ask only questions about technical uncertainties during the test, but not to communicate for other reasons so as not to disturb others. (2) The school representative compiled academic performance indicators: the results of diagnostic tests that were administered as a final assessment at the end of the 9th grade (primary school). Data were collected in an anonymized form, each student having their own code. The codes were assigned by the school based on the system recommended by the researcher (using letters + numbers denoting school, class and student number). The student was informed of his or her code shortly before the online testing, and then the student wrote this code on the testing site when starting the tests. The same code for each student was used when the academic achievement indicators were administered and sent to the researcher. Before the research started, each school had been informed about the aims and procedure of the research, and an informative letter to the parents was sent out by the school to allow for the participation in the study.

2.4. Data Analysis

To answer research questions, the following statistical analysis methods were used: multiple regression analysis, *t*-test and Spearman's correlation coefficients. The data were analyzed with statistics package SPSS version 22.

3. Results

First, the descriptive statistics of the indicators measured are presented (see Table 1). As we can see in Table 1, the amount of data regarding the results of the diagnostic test in English is not sufficient to perform further analyses.

Table 1. Descriptive statistics of the indicators measured in this study.

Measured Indicator	N	Min	Max	M	SD
Parental education level	630	1	6	3.44	1.24
Age of the student	655	14	17	15.41	0.53
I have felt difficulties dealing with studies during distance learning	659	0	5	3.11	1.32
The technological means available to me at home are sufficient to study remotely	659	0	5	4.36	0.97
Fluid nonverbal reasoning	534	0.00	10.00	4.96	2.63
Verbal reasoning	615	0.00	10.00	5.81	2.80
Self-management scale	647	1.00	30.00	16.31	6.11
Problem-solving: scale *Solution development and evaluation*	649	0.00	30.00	14.91	5.20
Problem-solving: scale *Flexibility to change the solution*	649	0.00	20.00	12.85	3.45
Diagnostic test in English	77	53.00	100.00	85.64	10.52
Diagnostic test in Latvian	330	14.29	99.09	61.17	15.99
Diagnostic test in Mathematics	347	10.67	100.00	61.44	22.25

The internal consistency measured by Cronbach's alpha was calculated for Self-management (α = 0.76), Problem-solving: scale Solution development and evaluation (α = 0.77) and Problem-solving: scale Flexibility to change the solution (α = 0.70), showing appropriate levels in the current sample. The data show that students come from classrooms with 9 to 34 students per class (M = 22.21; SD = 4.71), and such a variety is typical in Latvia, if students from smaller schools are compared to students from large schools.

Table 2 shows the Spearman's correlation coefficients between the measured indicators. The results of the diagnostic test in Latvian have a significant relation to Parental education level, Fluid nonverbal reasoning and Verbal reasoning, the Self-management scale and both scales of Problem solving: Solution development and evaluation and Flexibility to change the solution. The results of the diagnostic test in Mathematics show statistically significant correlations with Parental education level, Fluid nonverbal reasoning and Verbal reasoning, the Self-management scale and one scale of Problem-solving: Flexibility to change the solution, and are negatively correlated to the subjectively felt difficulties in dealing with distance learning. Both diagnostic tests (Mathematics and Latvian) show a significant interrelation, with r = 0.62. No significant correlation was found between the age of participants and the result of the diagnostic test in Mathematics (r = -0.05, p = 0.39) or the results of the diagnostic test in Latvian (r = -0.03, p = 0.64), and thus age would not be further included in the regression analysis. No significant difference was found between the gender of participants and the result of the diagnostic test in Mathematics the with statistical t-test analysis (t = 0.38, p = 0.71), with M = 61.81, SD = 22.65 for girls and M = 60.89, SD = 21.84 for boys. However statistically significant gender differences were found in the results of the diagnostic test in Latvian (t = 4.12, p = 0.00), with higher results for girls (M = 64.29, SD = 15.74) and lower results for boys (M = 57.15, SD = 15.43), indicating that gender should be included in the regression analysis.

Table 2. Correlations between the measured indicators for the students in the 9th grade.

			1	2	3	4	5	6	7	8	9	10
1.	Parental education level	Correlation Coefficient	1.00									
		n	630									
2.	I have felt difficulties dealing with studies during distance learning	Correlation Coefficient	−0.11 **	1.00								
		n	630	659								
3.	The technological means available to me at home are sufficient to study remotely	Correlation Coefficient	0.06	−0.05	1.00							
		n	630	659	659							
4.	Diagnostic test in Latvian	Correlation Coefficient	0.25 **	−0.09	−0.01	1.00						
		n	317	330	330	330						
5.	Diagnostic test in Mathematics	Correlation Coefficient	0.32 **	−0.13 *	0.04	0.62 **	1.00					
		n	332	347	347	330	347					
6.	Fluid nonverbal reasoning	Correlation Coefficient	0.16 **	−0.05	0.02	0.35 **	0.38 **	1.00				
		n	510	534	534	270	282	534				
7.	Verbal reasoning	Correlation Coefficient	0.17 **	−0.04	0.04	0.52 **	0.49 **	0.45 **	1.00			
		n	588	615	615	312	326	501	615			
8.	Self-management	Correlation Coefficient	0.08	−0.12 **	0.11 **	0.21 **	0.13 *	−0.023	0.012	1.00		
		n	619	647	647	329	345	526	609	647		
9.	Solution development and evaluation	Correlation Coefficient	0.09 *	−0.02	0.11 **	0.13*	0.06	−0.07	−0.01	0.45 **	1.00	
		n	621	649	649	329	345	528	611	647	649	
10.	Flexibility to change the solution	Correlation Coefficient	0.18 **	−0.01	0.13 **	0.34 **	0.25 **	0.22 **	0.25 **	0.37 **	0.46 **	1.00
		n	621	649	649	329	345	528	611	647	649	649

* $p < 0.05$; ** $p < 0.01$.

A multiple regression analysis was performed separately for the results of the diagnostic tests in Mathematics and Latvian to examine which were the best predictors. First, the regression analysis for the diagnostic test in Mathematics (DM) was performed. The indicators that correlate to the results in DM were included as independent variables in the hierarchical regression analysis (see Table 3). In the first step, the level of parents' education was entered, and explains 13% of the variation in the results of DM. Further, the problem-solving aspect Flexibility to change the solution was included and adds 4% to the variation. Self-management is not a statistically significant predictor of the results of DM. However, Fluid nonverbal reasoning and Verbal reasoning complement the additional 10% and 7%, respectively, to predict the variance in DM. The results indicate that higher students' Parental education level, Nonverbal reasoning, Verbal reasoning and Flexibility to change the solution led to higher results in the diagnostic test in Mathematics. It can be seen, however, that in Step 5, where all other measurements are included, both of the self-assessed measures (Self-management and Flexibility to change the solution) do not show statistically significant results ($\beta = 0.11$ and $\beta = -0.04$, respectively). When a simple regression was calculated, entering only the indicator Flexibility to change the solution as an independent variable, it showed that this indicator alone could explain 7% of the variation in DM ($R^2 = 0.07$, $F = 24.17$, $p = 0.000$; $B = 1.60$, $SE = 0.33$, $\beta = 0.26$, $p = 0.000$).

Table 3. Regression analysis of the result of the diagnostic test in Mathematics (n = 256) with independent variables: Parental education level, one scale of Problem-solving: Flexibility to change the solution, Self-management, Fluid nonverbal reasoning and Verbal reasoning.

	B	SE	β	F	R^2	ΔR^2
Diagnostic test result in Mathematics						
Step 1				38.52 **	0.13	0.13
Parental education	6.27	1.01	0.36 **			
Step 2				11.96 **	0.17	0.04
Parental education	5.67	1.00	0.33 **			
Flexibility to change the solution	1.21	0.35	0.20 **			
Step 3				0.05	0.17	0.00
Parental education	5.69	1.01	0.33 **			
Flexibility to change the solution	1.24	0.39	0.21 **			
Self-management	−0.05	0.22	−0.02			
Step 4				34.59 **	0.27	0.10
Parental education	5.50	0.95	0.32 **			
Flexibility to change the solution	0.80	0.37	0.13 *			
Self-management	−0.09	0.20	−0.03			
Fluid nonverbal reasoning	2.75	0.47	0.33 **			
Step 5				24.68 **	0.34	0.07
Parental education	4.87	0.92	0.28 **			
Flexibility to change the solution	0.66	0.35	0.11			
Self-management	−0.14	0.19	−0.04			
Fluid nonverbal reasoning	1.69	0.50	0.20 **			
Verbal reasoning	2.31	0.47	0.29 **			

* $p < 0.05$; ** $p < 0.01$.

Secondly, the indicators related to the results of the diagnostic test in Latvian (DL) were included as independent variables in the multiple regression analysis (see Table 4) to find out which are the best predictors of DL.

It can be seen in Table 4 that, in the first step, gender explains 4% of the variance in DL (higher for girls) and, in addition, parental education level explains another 7% of this variance. Together, they explain 11% of the variance. Further, the problem-solving aspects Solution development and evaluation and Flexibility to change add an extra 2% and 8%, respectively. As in the regressions performed for the DM, Self-management does not predict the results of DL in a statistically significant manner. In Step 6 and Step 7 Fluid nonverbal reasoning explains an additional 7%, and Verbal reasoning explains an additional 10% in the variance of DL. In Step 7 the indicators Self-management and Solution development and evaluation, as well as gender, are not statistically significant. When the following variables were entered as independent variables in a separately performed multiple regression analysis—Parental education level, Flexibility to change the solution (Problem-solving), Fluid nonverbal reasoning and Verbal reasoning—it was shown that, together, they could predict the DL test results and explained 36% of the variance (R^2 = 0.36, F = 34.05, p = 0.000).

Returning to the research questions stated in this study, it can be concluded that (1) the best predictors for the results in DM are Parental education, Fluid nonverbal reasoning and Verbal reasoning; (2) the best predictors for the results in DL are Parental education, Flexibility to change the solution (an aspect of problem solving), Fluid nonverbal reasoning and Verbal reasoning; (3) Self-management cannot significantly predict the results of DM or DL, although it correlates to the results of both DM and DL; (4) only one of the aspects of problem solving, Flexibility to change the solution, is predictive of the results in diagnostic tests.

Table 4. Regression analysis of the result of the diagnostic test in Latvian (n = 244) with independent variables: Gender, Parental education level, Solution development and evaluation and Flexibility to change the solution, Self-management, Fluid nonverbal reasoning, Verbal reasoning.

	B	SE	β	F	R^2	ΔR^2
Diagnostic test result in Latvian						
Step 1				9.50 **	0.04	0.04
Gender	−6.31	2.05	−0.19 **			
Step 2				20.39 **	0.11	0.08
Gender	−6.47	1.96	−0.20 **			
Parental education	3.45	0.76	0.27 **			
Step 3				4.18 *	0.13	0.02
Gender	−6.04	1.97	−0.19 **			
Parental education	3.27	0.76	0.26 **			
Solution development and evaluation (Problem solving)	0.36	0.18	0.13 *			
Step 4				24.03 **	0.21	0.08
Gender	−3.79	1.94	−0.12			
Parental education	2.77	0.74	0.22 **			
Solution development and evaluation (Problem-solving)	−0.11	0.19	−0.04			
Flexibility to change the solution (Problem solving)	1.53	0.31	0.34 **			
Step 5				0.98	0.21	0.00
Gender	−3.33	1.99	−0.10			
Parental education	2.70	0.74	0.21 **			
Solution development and evaluation (Problem solving)	−0.18	0.21	−0.06			
Flexibility to change the solution (Problem solving)	1.47	0.32	0.33 **			
Self-management	0.17	0.17	0.07			
Step 6				23.11 **	0.28	0.07
Gender	−3.43	1.90	−0.11			
Parental education	2.56	0.71	0.20 **			
Solution development and evaluation (Problem solving)	0.06	0.20	0.02			
Flexibility to change the solution (Problem solving)	1.01	0.33	0.23 **			
Self-management	0.09	0.16	0.04			
Fluid nonverbal reasoning	1.72	0.36	0.28 **			
Step 7				38.14 **	0.38	0.10
Gender	−2.96	1.77	−0.09			
Parental education	1.90	0.67	0.15 **			
Solution development and evaluation (Problem solving)	0.12	0.19	0.04			
Flexibility to change the solution (Problem solving)	0.87	0.30	0.20 **			
Self-management	0.04	0.15	0.02			
Fluid nonverbal reasoning	0.76	0.37	0.13 *			
Verbal reasoning	2.08	0.34	0.37 **			

* $p < 0.05$; ** $p < 0.01$.

4. Discussion

One of the aims of the current research was to determine the best predictors of results in school diagnostic tests at the end of the 9th grade (considered as important indicators of academic achievement) during the distance learning due to the COVID-19 pandemic. The pandemic was an unprecedented problem, during which distance learning was introduced for students who had never learned in such a way. For many students, it was a new situation, posing many new problems (e.g., planning one's time, motivating oneself and lack of regime) (Hacatrjana 2021a). Thus, it was assumed that problem solving and self-management skills would be necessary to effectively learn independently and reach academic goals during this time, in parallel to cognitive abilities and parental education level, that have both proved to be important predictors of academic achievement (e.g., Flores-Mendoza et al. 2021).

The results of the current study show that there are some differences regarding the predictors of the results of diagnostic tests in different fields of study—Mathematics and Latvian. The best predictors for the diagnostic tests in Mathematics of 9th graders are their cognitive abilities (in this case—fluid nonverbal reasoning and verbal reasoning) and parental

education level, explaining altogether about a third of the variance in the Mathematics test. Only one aspect of problem solving—the flexibility to change the solution—showed an additional contribution that was statistically significant, when analyzed separately. As to the results of the diagnostic test in Latvian, the level of parents' education, the flexibility to change the solution (one aspect of problem-solving), fluid nonverbal reasoning and verbal reasoning have a predictive value. Together, these variables can explain more than a third of the variance in the results of the diagnostic test in Latvian.

A conclusion which can be generalized to the tests in both subjects (Latvian and Mathematics) is that a more important role is played by cognitive abilities (in this case fluid nonverbal and verbal abilities) in comparison to self-assessed indicators of skills. It is argued that students might give socially desirable answers to self-report questions or might not be precise enough in evaluating their abilities. Moreover, previous studies have shown that cognitive abilities assessed with tests are indeed the strongest predictor of academic performance (e.g., Demetriou et al. 2019; Frey 2019; Kampa et al. 2021), though the contribution is lower for older students compared to younger students. Conway and Hao (2020) argue for the need for precise methodologies if we want to assess the relation between non-cognitive factors and SAT scores. The authors argue that cognitive test scores typically explain at least half of the variation in SAT tests, if cognitive measurements have been adequately selected and cover a full range of abilities. In the current study the cognitive abilities did not explain such a large proportion of the variance, possibly due to this very reason.

The results presented here also showed the importance of parental education level to the school test results. Having parents with a higher level of education predicts higher results in academic achievement for students, and, as other research shows, it might be even more important during the pandemic (Easterbrook 2021). The tight relation between parental education and academic achievement is already established in previous studies and discussed in the literature (e.g., Idris et al. 2020; O'Leary and Marks 2021). It might be explained not only by the level of abilities, but also by higher parents' involvement and valuing education as important in life based on their own experience (Lara and Saracostti 2019). During distance learning, parents' involvement might have played an even larger role, and research shows that parental knowledge and comprehension of education, as well as proficiency in technology, was related to several indicators, such as the encouragement of an effective use of technology for education (Dimopoulos et al. 2021).

The second aim of the study was to examine the extent to which self-assessed problem-solving skills and self-management skills could predict the results in diagnostic tests of 9th graders during distance learning. Two aspects of problem solving were measured: the flexibility to change the solution and solution development and evaluation, and a total score of self-management was obtained. Compared to cognitive abilities and parental education level, these skills have a much smaller influence on the test results. Nevertheless, one aspect of problem solving in particular—the flexibility to change the solution—can explain a relatively small but statistically significant proportion of the variance of the test results. This aspect of problem-solving is briefly discussed below.

The flexibility to change the solution is an important aspect of problem solving (Hacatrjana 2021b) and was significantly predictive of the results in the diagnostic test in Latvian. It was also predictive of the diagnostic test results in Mathematics when analyzed separately. But when other variables are included into the regression, the significance of this indicator drops, and other variables—nonverbal and verbal reasoning, as well as parental education—become the most important predictors. Why is the flexibility to change the solution important to get better results in tests and why, in the current research, does it turn out to be more important than the ability to come up with solutions and evaluate them? The flexibility to change the solution might be a crucial aspect to successfully solve problems or tasks, providing that an individual is able to, first, detect if something is wrong in the solution; secondly, make a decision to start over or change something in the solution; and third, come up with an alternative or a new way to do the task and execute it. It might

be related to the ability to switch between ideas and possible solutions, and not to get stuck on the first solution that has come to mind. In the mathematics, the term "flexibility" characterizes the ability to choose between several solving options (meaning that the student is aware of various approaches and is able to implement them when necessary) and is also related to academic achievement (Hästö et al. 2019). The results of the current study might also be explained by the fact that the Flexibility to change is significantly correlated to both cognitive tasks: Fluid nonverbal reasoning and Verbal reasoning, while the other aspect of problem solving—Solution development and evaluation—shows weaker correlations with these tasks. The fact that the Solution development and evaluation aspect is less related to the diagnostic test results, compared to the Flexibility to change the solution, is worth studying further, to examine if the flexibility in one's actions during a problem- or task-solving process is crucial to successful problem solving in general, as these results suggest, and to what extent flexibility is related to cognitive abilities and might be taught as a skill and an attitude.

Another important finding in the current study is that the self-management skills failed to show statistically significant results in regression analysis to predict the results in the Mathematics and Latvian tests, though self-management skills were correlated significantly to the results of these tests. This contradicts previous research showing a significant relation between self-management or other aspects of self-regulated learning and indicators and academic achievement (Pintrich et al. 1993; Zimmerman and Martinez-Pons 1988; Abd-El-Fattah 2010; Veenman et al. 2014), and the results from a study where students revealed the importance of self-management skills during distance learning (Hacatrjana 2021a). How could these results be explained, considering the importance of self-management (and self-regulated learning as a broader term) in education? One of the explanations is that these studies vary in the methods used and the conceptualization of terms, such as self-management in learning or self-regulated learning. Another explanation is that other indicators measured in this study are just stronger predictors, having tighter correlations to the test results, and thus statistically self-management skills are left below the line. This could also be due to the conceptualization of the self-management construct that was measured in the current study. It covers the actions of planning and organizing one's learning process and physical settings and maintaining the motivation to do the school tasks but does not cover broader aspects of self-regulated learning, such as implementing learning strategies. One explanation for these results might be that the skills included in this concept are indeed important for focusing on the studies and an accurate approach to learning on a daily basis and managing daily learning tasks, but they are not sufficient to increase the level of performance in the academic tests.

Overall, the currently presented results reveal that a relatively little contribution is made by the problem-solving and self-management skills, assessed with self-assessment methods, to the results of school tests during the pandemic. Yet, an important aspect of problem-solving skills—flexibility to change the solution—does make an additional contribution and explain the variance of the test results, especially in the Latvian test. There are no comparison data on students' problem-solving or self-management skills before the pandemic. Nevertheless, the importance of developing students' skills and habits to effectively deal with problems and obstacles should not be neglected, as it is previously proven that some aspects of problem solving can be successfully developed in the classroom (Verschaffel et al. 1999). While reasoning abilities are crucial for doing well in the diagnostic tests, it is also important to teach students the skills and strategies to apply when facing new or complicated tasks, so that they can think of solutions, implement them, and make the decision to change something in the solution if it turns out to be inappropriate.

Limitations

For some measurements (scores in the diagnostic tests), the data were not fully provided by the schools, mostly due to lack of capacity of workforce resources (the data needed to be coded to anonymize students' names). This led to a smaller amount of data used for

such measurements as diagnostic tasks. The study would have benefited if a wider variety of cognitive measures had been used to examine if an even larger contribution would be made to explaining the variance in students' test results. The level of students' assessment of their skills before the pandemic is not known, and thus a conclusion on the dynamic of these skills before and during the pandemic cannot be drawn—only conclusions on the relation between skills and academic achievement during the pandemic. The study involved the self-assessed measurements that were previously discussed, and it would have benefited if the teachers' ratings of students' skills (for example, self-management) had been used to enhance the validity of these measurements. However, during distance learning, the teachers could not directly observe how a student is organizing his or her daily learning process, and the ratings would also be based on their previous experience.

5. Conclusions

Several indicators that might have predicted the results of diagnostic tests in Mathematics and Latvian at the end of primary school (9th grade) during the unprecedented COVID-19 pandemic and the corresponding distance learning were analyzed in this study: cognitive abilities, verbal reasoning and nonverbal reasoning, self-reported problem-solving skills and self-management skills and parental education level. The most important predictors for the test results were cognitive abilities and parental education level, and only one aspect of problem solving: the flexibility to change the solution. Self-assessed problem-solving skills and self-management skills did not play such an important role in predicting the results in the diagnostic tests taken during distance learning, as would be expected. We can speculate that self-management skills were probably important in the daily management of one's learning during distance learning, as shown by previous research, but they were not decisive to reach higher academic results in tests at the end of the 9th grade. One aspect of problem-solving—the flexibility to change the solution—contributed to the results in diagnostic tests, especially in Latvian (the native language), indicating that this might be an important set of skills and attitudes for students to develop to successfully deal with school tasks. Based on this analysis, conclusions can be drawn about the importance of these skills to maintain academic achievement when students are facing new situations. It can be further assumed that it is justified to teach problem-solving skills as part of the curriculum, as they might help students adapt to other unprecedented events in the future.

Funding: This research was funded by the European Regional Development Fund under the activity "Post-doctoral Research Aid" project "Relationship between students' self-management and problem-solving skills and changes in academic achievement during face to face and distance learning situations", No. 1.1.1.2/VIAA/4/20/697.

Institutional Review Board Statement: Data were collected in accordance with the rules of the hosting University and the Declaration of Helsinki.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study (information about the research was also sent to the parents by the participating schools with an option to withdraw their child's participation).

Data Availability Statement: The data are available from the author by request.

Conflicts of Interest: The author declares no conflict of interest.

References

Abd-El-Fattah, Sabry. 2010. Garrison's Model of Self-Directed Learning: Preliminary Validation and Relationship to Academic Achievement. *The Spanish Journal of Psychology* 13: 586–96. [CrossRef] [PubMed]

Andrade, Heidi L. 2019. A Critical Review of Research on Student Self-Assessment. *Frontiers in Education* 4: 87. [CrossRef]

Brown, Gavin T. L., and Lois R. Harris. 2013. Student self-assessment. In *The SAGE Handbook of Research on Classroom Assessment*. Edited by James H. McMillan. Thousand Oaks: Sage, pp. 367–93.

Cabinet of Ministers Republic of Latvia. 2018. Rules Nr.747 from 27.11.2018. Available online: https://likumi.lv/ta/id/303768 (accessed on 12 October 2021).

Cabinet of Ministers Republic of Latvia. 2021. Rules Nr.319, (Chapter of MK Rule Nr. 190, redaction of 24.03.2021). Available online: https://likumi.lv/ta/id/315040-noteikumi-par-valsts-parbaudes-darbu-norises-laiku-2020-2021-nbsp-macibu-gada (accessed on 4 October 2021).

Carter, Richard A., Jr., Mary Rice, Sohyun Yang, and Haidee A. Jackson. 2020. Self-regulated learning in online learning environments: Strategies for remote learning. *Information and Learning Sciences* 121: 321–29. [CrossRef]

Chuderski, Adam, and Jan Jastrzębski. 2018. Much ado about aha!: Insight problem solving is strongly related to working memory capacity and reasoning ability. *Journal of Expermental Psychology: General* 147: 257–81. [CrossRef] [PubMed]

Conway, Andrew R. A., and Han Hao. 2020. The Role of Non-Cognitive Factors in the SAT Remains Unclear: A Commentary on Hannon (2019). *Journal of Intelligence* 8: 15. [CrossRef]

Demetriou, Andreas, Smaragda Kazi, George Spanoudis, and Nikolaos Makris. 2019. Predicting school performance from cognitive ability, self-representation, and personality from primary school to senior high school. *Intelligence* 76: 101381. [CrossRef]

Dimopoulos, Kostas, Christos Koutsampelas, and Anna Tsatsaroni. 2021. Home schooling through online teaching in the era of COVID-19: Exploring the role of home-related factors that deepen educational inequalities across European societies. *European Educational Research Journal* 20: 479–97. [CrossRef]

Doyle, Audrey, Zita Lysaght, and Michael O'Leary. 2021. High stakes assessment policy implementation in the time of COVID-19: The case of calculated grades in Ireland. *Irish Educational Studies* 40: 385–98. [CrossRef]

Easterbrook, Matthew J. 2021. Inequalities in How Difficult Pupils Find it to Complete Their School Work from Home, and Why That Might Be. Report. Available online: https://www.inpsyed.net/post/inequalities-in-home-learning-difficulty-to-complete-tasks-and-reasons-why (accessed on 30 November 2021).

Ellis, Derek M., Matthew K. Robison, and Gene A. Brewer. 2021. The Cognitive Underpinnings of Multiply-Constrained Problem Solving. *Journal of Intelligence* 9: 7. [CrossRef]

Engzell, Per, Arun Frey, and Mark D. Verhagen. 2021. Learning loss due to school closures during the COVID-19 pandemic. *Proceedings of the National Academy of Sciences USA* 118: e2022376118. [CrossRef]

Fischer, Andreas, Samuel Greiff, and Joachim Funke. 2012. The Process of Solving Complex Problems. *The Journal of Problem Solving* 4: 19–42. [CrossRef]

Flores-Mendoza, Carmen, Ruben Ardila, Miguel Gallegos, and Norma Reategui-Colareta. 2021. General Intelligence and Socioeconomic Status as Strong Predictors of Student Performance in Latin American Schools: Evidence From PISA Items. *Frontiers in Education* 6: 632289. [CrossRef]

Frensch, Peter A., and Joachim Funke. 1995. Definitions, Traditions and a General Framework for Understanding Complex Problem Solving. In *Complex Problem Solving: The European Perspective*. Edited by P. A. Frensch and J. Funke. New York: Lawrence Erlbaum Associates, pp. 3–26.

Frey, Meredith C. 2019. What We Know, Are Still Getting Wrong, and Have Yet to Learn about the Relationships among the SAT, Intelligence and Achievement. *Journal of Intelligence* 7: 26. [CrossRef]

Furnham, Adrian, and Simmy Grover. 2020. Correlates of Self-Estimated Intelligence. *Journal of Intelligence* 8: 6. [CrossRef]

Garrisson, Donn R. 1997. Self-Directed Learning: Toward a Comprehensive Model. *Adult Education Quarterly* 48: 18–33. [CrossRef]

González-Betancor, Sara M., Alicia Bolívar-Cruz, and Domingo Verano-Tacoronte. 2019. Self-assessment accuracy in higher education: The influence of gender and performance of university students. *Active Learning in Higher Education* 20: 101–14. [CrossRef]

Greiff, Samuel, Sascha Wüstenberg, Gyöngyvér Molnár, Andreas Fischer, Joachim Funke, and Benó Csapó. 2013. Complex problem solving in educational contexts—Something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology* 105: 364–79. [CrossRef]

Greiff, Samuel, Andre Kretzschmar, Jonas C. Müller, Birgit Spinath, and Romain Martin. 2014. The Computer-Based Assessment of Complex Problem Solving and How It Is Influenced by Students' Information and Communication Technology Literacy. *Journal of Educational Psychology* 106: 666–80. [CrossRef]

Hacatrjana, Liena. 2021a. Ability to deal with it: Self-management and problem-solving skills, motivation and routines helped high-school students during COVID-19 pandemic. In *Human, Technologies and Quality of Education, 2021*. Proceedings of Scientific Papers. Riga: University of Latvia Press, pp. 126–36. [CrossRef]

Hacatrjana, Liena. 2021b. Assessment of students' problem-solving skills and self-management skills: Two new questionnaires for assessment. Paper presented at conference "The World of Didactics: Didactics in the Contemporary World", Kyiv, Ukraine, September 21–22. Available online: https://sites.google.com/view/conferencedidactica2021/%D0%B7%D0%B1%D1%96%D1%80%D0%BD%D0%B8%D0%BA-%D0%BC%D0%B0%D1%82%D0%B5%D1%80%D1%96%D0%B0%D0%BB%D1%96%D0%B2-%D0%BA%D0%BE%D0%BD%D1%84%D0%B5%D1%80%D0%B5%D0%BD%D1%86%D1%96%D1%97?authuser=0 (accessed on 4 November 2021).

Harris, Alexandra. M., Jeremiah T. McMillan, Benjamin Listyg, Laura E. Matzen, and Nathan Carter. 2020. Measuring Intelligence with the Sandia Matrices: Psychometric Review and Recommendations for Free Raven-Like Item Sets. *Personnel Assessment and Decisions* 6: 3. [CrossRef]

Hästö, Peter, Riikka Palkki, Dimitri Tuomela, and Jon R. Star. 2019. Relationship between flexibility and success in national examinations. *European Journal of Science and Mathematics Education* 7: 1–13. [CrossRef]

Heppner, Paul P., and Chris H. Petersen. 1982. The development and implications of a personal problem solving inventory. *Journal of Counseling Psychology* 29: 66–75. [CrossRef]

Idris, Muhammad, Sajjad Hussain, and Ahmad Nasir. 2020. Relationship between Parents' Education and their children's Academic Achievement. *Journal of Arts and Social Sciences* VII: 82–92. [CrossRef]

Kampa, Nele, Ronny Scherer, Steffani Saß, and Stefan Schipolowski. 2021. The relation between science achievement and general cognitive abilities in large-scale assessments. *Intelligence* 86: 101529. [CrossRef]

Kim, Young-Hoon, Chi-yue Chiu, and Zhimin Zou. 2010. Know thyself: Misperceptions of actual performance undermine achievement motivation, future performance, and subjective well-being. *Journal of Personality and Social Psychology* 99: 395–409. [CrossRef] [PubMed]

Kretzschmar, Andre, Liena Hacatrjana, and Malgozata Rascevska. 2017. Re-evaluating the Psychometric Properties of MicroFIN: A Multidimensional Measurement of Complex Problem Solving or a Unidimensional Reasoning Test? *Psychological Test and Assessment Modeling* 59: 157–82.

Kuhfeld, Megan, Karyn Lewis, Patrick Meyer, and Beth Tarasawa. 2020. *Comparability Analysis of Remote and in-Person MAP Growth Testing in Fall 2020*. NWEA. Available online: https://www.nwea.org/research/publication/comparability-analysis-of-remote-and-in-person-map-growth-testing-in-fall-2020/ (accessed on 29 October 2021).

Kvedere, Liene. 2014. Mathematics Self-efficacy, Self-concept and Anxiety Among 9th Grade Students in Latvia. *Procedia Social and Behavioral Sciences* 116: 2687–90. [CrossRef]

Lara, Laura, and Mahia Saracostti. 2019. Effect of Parental Involvement on Children's Academic Achievement in Chile. *Frontiers in Psychology* 10: 1464. [CrossRef]

Matzen, Laura E., Zachary O. Benz, Kevin R. Dixon, Jamie Posey, James K. Kroger, and Ann E. Speed. 2010. Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior Research Methods* 42: 525–41. [CrossRef]

Miller, Tyler M., and Lisa Geraci. 2011. Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37: 502–6. [CrossRef]

Ministry of Education and Science of Latvia. 2020. *Attālinātās Mācības no Skolēnu, Skolotāju un Vecāku Skatupunkta*; [Distance Learning from the Viewpoint of Students, Teachers and Parents]. Published on July 14. Available online: https://www.izm.gov.lv/lv/jaunums/attalinatas-macibas-no-skolenu-skolotaju-un-vecaku-skatupunkta-1 (accessed on 20 October 2021).

Nota, Laura, Paul P. Heppner, Salvatore Soresi, and Mary J. Heppner. 2009. Examining Cultural Validity of the Problem-Solving Inventory (PSI) in Italy. *Journal of Career Assessment* 17: 478–94. [CrossRef]

O'Leary, Michael, and Gary N. Marks. 2021. Are the effects of intelligence on student achievement and well-being largely functions of family income and social class? Evidence from a longitudinal study of Irish adolescents. *Intelligence* 84: 101511. [CrossRef]

OECD. 2013. *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD Publishing. [CrossRef]

OECD. 2017. *PISA 2015 Results (Volume V): Collaborative Problem Solving*. Paris: OECD Publishing. [CrossRef]

OECD. 2020. *Lessons for Education from COVID-19: A Policy Maker's Handbook for More Resilient Systems*. Paris: OECD Publishing, Available online: http://www.oecd.org/education/lessons-for-education-from-covid-19-0a530888-en.htm (accessed on 8 October 2021).

Pintrich, Paul R., David A. F. Smith, Teresa Garcia, and Wilbert J. Mckeachie. 1993. Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement* 53: 801–13. [CrossRef]

Polya, George. 1957. *How to Solve It. A New Aspect of Mathematical Method*, 2nd ed. Princeton: Princeton University Press.

Rammstedt, Beatrice, and Thomas H. Rammsayer. 2002. Self-estimated intelligence: Gender differences, relationship to psychometric intelligence and moderating effects of level of education. *European Psychologist* 7: 275–84. [CrossRef]

Reimers, Fernando M., and Andreas Schleicher. 2020. *A Framework to Guide an Education Response to the COVID-19 Pandemic of 2020*. Paris: OECD, Available online: https://globaled.gse.harvard.edu/files/geii/files/framework_guide_v1_002.pdf (accessed on 2 September 2021).

Rogers, Adam A., Thao Ha, and Sydney Ockey. 2021. Adolescents' Perceived Socio-Emotional Impact of COVID-19 and Implications for Mental Health: Results From a U.S.-Based Mixed-Methods Study. *Journal of Adolescent Health* 68: 43–52. [CrossRef]

Rosen, Maya L., Alexandra M. Rodman, Steven W. Kasparek, Makeda Mayes, Malila M. Freeman, Liliana J. Lengua, Andrew N. Meltzoff, and Katie A. McLaughlin. 2021. Promoting Youth Mental Health during COVID-19: A Longitudinal Study Spanning Pre- and Post-Pandemic. *PLoS ONE* 16: e0255294. [CrossRef]

Sanchez, Carmen E., Kayla M. Atkinson, Alison C. Koenka, Hannah Moshontz, and Harris Cooper. 2017. Self-Grading and Peer-Grading for Formative and Summative Assessments in 3rd Through 12th Grade Classrooms: A Meta-Analysis. *Journal of Educational Psychology* 109: 1049–66. [CrossRef]

Scott, Samantha R., Kenia A. Rivera, Ella Rushing, Erika M. Manczak, Christopher S. Rozek, and Jenalee R. Doom. 2021. "I Hate This": A Qualitative Analysis of Adolescents' Self-Reported Challenges During the COVID-19 Pandemic. *Journal of Adolescent Health* 68: 262–269. [CrossRef]

Tekkol, Ilkay A., and Melek Demirel. 2018. An Investigation of Self-Directed Learning Skills of Undergraduate Students. *Frontiers in Psychology* 9: 2324. [CrossRef]

Thorn, William, and Sthepan Vincent-Lancrin. 2021. *Schooling During a Pandemic: The Experience and Outcomes of Schoolchildren During the First Round of COVID-19 Lockdowns*. Paris: OECD Publishing. [CrossRef]

Vasileiadou, Despina, and Konstantinos Karadimitriou. 2021. Examining the impact of self-assessment with the use of rubrics on primary school students' performance. *International Journal of Educational Research Open* 2: 100031. [CrossRef]

Veenman, Marcel V. J., Rob D. Hesselink, Shannon Sleeuwaegen, Sophie I. E. Liem, and Marieke G. P. Van Haaren. 2014. Assessing Developmental Differences in Metacognitive Skills With Computer Logfiles: Gender by Age Interactions. *Psychological Topics* 23: 99–113.

Verschaffel, Lieven, and Erik De Corte. 1993. A decade of research on word problem solving in Leuven: Theoretical, methodological, and practical outcomes. *Education Psychology Review* 5: 239–56. [CrossRef]

Verschaffel, Lieven, Erik De Corte, Sabien Lasure, Griet Van Vaerenbergh, Hedwig Bogaerts, and Elie Ratinckx. 1999. Learning to Solve Mathematical Application Problems: A Design Experiment With Fifth Graders. *Mathematical Thinking and Learning* 1: 195–229. [CrossRef]

Zimmerman, Barry J. 2008. Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal* 45: 166–83. [CrossRef]

Zimmerman, Barry J., and Manuel Martinez-Pons. 1988. Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology* 80: 284–290. [CrossRef]

Journal of
Intelligence

MDPI

Article

Assessing Children 'At Risk': Translation and Cross-Cultural Adaptation of the Motor Behavior Checklist (MBC) into Arabic and Pilot Use in the United Arab Emirates (UAE)

Maria Efstratopoulou [1], Hala Elhoweris [1], Abeer Arafa Eldib [1] and Eleni Bonti [2,3,*]

[1] Department of Special Education (CEDU), United Arab Emirates University (UAEU), Al-Ain 112612, United Arab Emirates; maria.efstratopoulou@uaeu.ac.ae (M.E.); halae@uaeu.ac.ae (H.E.); eldib@ese.gov.ae (A.A.E.)
[2] First Psychiatric Clinic, "Papageorgiou" General Hospital, School of Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki, Ring Road Thessaloniki, N. Efkarpia, 54603 Thessaloniki, Greece
[3] Department of Education, School of Education, University of Nicosia, Nicosia 2417, Cyprus
* Correspondence: bonti@auth.gr

Abstract: Children's emotional, behavioral, and developmental problems can be properly identified and assessed based on observations from their teachers and parents. The Motor Behavior Checklist (MBC) was designed to assist classroom teachers and Physical Education (PE) teachers in assessing their students' motor-related behaviors. The instrument has already been successfully translated and culturally adapted into six languages and used in a number of research studies internationally. The present study aimed to develop the Arabic version of the MBC checklist and proceed with the necessary cross-cultural adaptations for the use of the instrument in Arabic speaking countries and especially in United Arab Emirates (UAE) primary schools. The translation and cultural adaptation of the MBC was based on the ten-step process: forward translation of the original instrument; development of a synthesized version, back-translation; linguistic and semantic comparisons; back translators evaluation of divergent items; development of a synthesized version; based on the back translators' suggestions; clarity assessment of the synthesized version by professionals (teachers); additional assessment of clarity indicators by a focus group of experts; and development of the final version. Results indicated a satisfactory level of agreement between the original and the back-translated versions, while nine items required minor adjustments and two items needed major adaptations and word replacements to clarify their content and be fully adapted into the UAE culture. In the pilot use, UAE teachers confirmed the clarity of the items in an 84% percentage. The final translated version's overall content was found sufficiently compatible with the original version of the instrument. The study highlights the importance of a rigorous translation process and the process of cultural adaptation.

Keywords: Motor Behavior Checklist (MBC); cross-cultural adaptation; UAE; physical education; behavioral problems

Citation: Efstratopoulou, Maria, Hala Elhoweris, Abeer Arafa Eldib, and Eleni Bonti. 2022. Assessing Children 'At Risk': Translation and Cross-Cultural Adaptation of the Motor Behavior Checklist (MBC) into Arabic and Pilot Use in the United Arab Emirates (UAE). *Journal of Intelligence* 10: 11. https://doi.org/10.3390/jintelligence10010011

Received: 11 November 2021
Accepted: 31 January 2022
Published: 5 February 2022

1. Introduction

1.1. The Manifestation of NDD in School-Aged Children

Neurodevelopmental disorders (NDD) constitute a group of disorders that commonly emerge during childhood or adolescence and usually affect behaviors that are significant for normal interactions, ranging from school to social occasions (American Psychiatric Association 2013). NDD include Intellectual Disability Communication Disorders, Autism Spectrum Disorder (ASD), Attention-Deficit/Hyperactivity Disorder (ADHD), Specific Learning Disorder (SLD), and Motor Disorders. They manifest early in development, often before school entry age, affecting the child's personal, academic, and social functioning. The range of NDD developmental deficits may vary from very specific limitations of learning

or control of executive functions to global impairments of intelligence or social skills (American Psychiatric Association 2013). Research indicates that Neurodevelopmental Disorders in children may co-occur and there are some common behavioral, cognitive, andemotional characteristics, which often lead to a challenging or invalid diagnosis or a misdiagnosis (Jensen et al. 2001). It is often the case that many children with NDD enter elementary school without having received a formal diagnosis. As a result, their cognitive/learning, language, attentive, emotional, and/or behavioral difficulties (EBP) often lead them to 'school failure' along with low self-esteem issues, while in certain cases, they may face more severe psychological or psychiatric conditions as adults (Eisenberg et al. 2000).In addition, unlike learning or cognitive assessment, which can be attained through direct assessment, children's emotional and behavioral competencies can only be properly identified and described mainly based on the reports of others (Efstratopoulou et al. 2015).

1.2. Students of Determination in United Arabs Emirates

The number of pupils with a disability enrolled in the emirate's schools increased by more than 3500 in the past year, the Knowledge and Human Development Authority revealed. The results of the 2018–2019 inspections by Dubai's private schools' regulator revealed that pupils with physical and intellectual disability account for six per cent of the 278,794-strong school population. Inspections also revealed that 71 per cent of schools provide a good or better quality of education for disabled pupils, also referred to as 'people of determination' by the UAE government.

Early assessment is one of the most important components in the UAE education system according to the new National Standards for Education presented by the Ministry of Education. The new strategy of the Ministry of Education (MOE) focuses on supporting early assessment and professional development and training for teachers. In this direction, the translation and cultural adaptations of valid assessments tools (like the MBC for children) that can be used by professionals (teachers/physical educators) in schools settings is of high importance.

1.3. Teachers and Parents' Involvement in the Assessment Process in Children

Researchers have recognized the importance of parents and teachers' roles in obtaining a more holistic and valid assessment of children's emotional and behavioral functioning. Especially, as regards the evaluation of EBP encountered by children with ADHD (e.g., inattention, lack of concentration, impulsivity, hyperactivity, learning problems, etc.), teachers and parents are often considered as the principal agents of the assessment process (Gardon 2012). More specifically, with the use of several assessment scales, they are asked to rate their students/children's behaviors across a variety of settings (e.g., home, school, athletic activities, play, etc.) in common, everyday circumstances (Paiano et al. 2019). These behaviors are often related to problems in academic achievement, social maturity issues, and family relationships, which are usually evident in children with NDD (Haack et al. 2019).Furthermore, teachers, in particular, are considered as the most representative and valid source of information regarding their students' behavioral problems, since they interact with the children on a daily basis and within the school environment, where specific behavioral patterns are expected (Gardon 2012).

1.4. PE Teachers' Role in the Identification of Emotional and Behaviour Problems in Children (EBP)

A significant body of research has indicated that Physical Education (PE) teachers, in particular, are among the most appropriate educators in the assessment process, since they have the knowledge and skills to better identify the warning signs of abnormal, especially motor-related behaviors. Therefore, they can provide valuable information about the overall development and EBP manifestation of their students (Kashani et al. 1997; Mol Lous et al. 2002). More precisely, as Kashani et al. (1997) have argued, evidence for the presence of externalizing and/or internalizing behavioral symptoms can be better obtained in multiple active situations during PE classes and team games. In addition,

specific problematic behaviors (e.g., attention, contact, learning, or mood difficulties) can be systematically observed during standardized play procedures (Mol Lous et al. 2002). Hence, PE lessons and group play situations offer multiple opportunities for observing a child while s/he engages with his/her classmates in various settings (e.g., inside or outside the classroom, at the gym, during group activities or games) (Efstratopoulou et al. 2015).

1.5. School-Based Behavioural Evaluation Instruments

The majority of school-based behavioral assessment tools are structured for use by school psychologists and/or counsellors, often use mental health terminologies, are time-consuming, and are not intended for use by teachers or physical educators in school settings. Data obtained through such rating scales are interpreted mainly by medical professionals and might lead to an official diagnosis. Thus, psychotherapeutic treatments and/or psychiatric therapy usually follow such diagnoses. However, teachers need behavioral assessment tools that are easy to use and interpret, to monitor the effectiveness of their supporting strategies and educational approaches in-class or school-based behavior support plans. School-based behavior rating scales must be valid, reliable, and concise screening tools, which lead to the early identification of children 'at risk' of developing behavioral, NDD, or other disorders. In addition, they should be available for use by classroom teachers and school staff and should be integrated into the mainstream educational system regularly (i.e., they could be administered to all students during their first years of primary education). Furthermore, such tools should also serve as a means for additional, more detailed assessments of students who meet the 'at risk' criteria, to safeguard that they will eventually receive the appropriate intervention at an early stage, thus preventing secondary problems, which usually accompany NDD, as the child grows older. Finally, they might also serve as a measure of the effectiveness of the intervention programs designed for these children, through the pre-test (prior intervention) and post-test (following intervention) result analysis process (Efstratopoulou et al. 2015).

1.6. The Motor Behaviour Checklist (MBC)

The Motor Behavior Checklist (MBC; Efstratopoulou et al. 2012b) is a practical, easy to administer, useful, and valid measure for observing the motor behavior of children aged between 6 and12 years, and for screening and assessing children with EBP problems and possible underlying disorders (e.g., motor-development problems, Autism Spectrum disorders, ADHD, learning difficulties, etc.) in the school environment. Its first version was standardized in the British primary school-age population (initially including 150 items). In its final version, it includes 59 items describing observable 'problematic' behaviors. The higher the total score recorded the more indicative of possible underlying problems. The instrument can provide separate scores for each of the seven factors and total externalizing/internalizing behavior scores. Externalizing behaviors include rule-breaking (seven items), hyperactivity and impulsivity (14 items), and lack of attention (10 items). Internalizing behaviors include low energy (four items), stereotyped behaviors (two items), lack of social interaction (10 items), and lack of self-regulation (12 items).

The MBC can be completed by the class teacher or the physical educator (provided that he/she knows the child well enough (i.e., more than six months)), using a 5-point Likert scale ranging from never (0) to almost always (4) indicating the frequency of the observed behavior (Efstratopoulou et al. 2019). The MBC does not require the physical presence of the child during administration, and data can be collected even via online procedures. The checklist cannot be solely used as a diagnostic tool but can be part of a broader assessment battery and is a quick and cost-effective screening tool/measure for the early identification of emotional–behavioral symptoms/difficulties in typical school-aged children, who are often left unnoticed because they are underdiagnosed or misdiagnosed. Finally, administration and completion of the MBC checklist do not require verbal skills on the child's part and can provide a detailed individual profile on different areas of the

child's development (e.g., social skills, self-regulation, aggressiveness, hyperactivity, etc.), while assessing deviant behaviors in school settings (Efstratopoulou et al. 2015).

1.7. Psychometric Properties of MBC

Research on the psychometric properties of the instrument indicates that the MBC for children is a content-homogeneous instrument that has a strong correlation and is highly stable (Efstratopoulou et al. 2012a). More specifically, the internal consistency coefficients ranged from 0.82 to 0.95, reproducibility was 0.85 to 0.90 according to intraclass correlation coefficients (ICC), and concordance was also 0.75 to 0.91. Discriminant and convergent validity of the MBC scale were established, and it has been proven that MBC can provide valid ratings on externalizing and internalizing behaviors of school-aged children (Efstratopoulou et al. 2012b).

Although the MBC for children is not a diagnostic tool itself, it can provide valid complementary information on attentional, emotional, and developmental problems in children when used by physical educators in school settings. It is a new practical and useful measure for assessing externalizing and/or internalizing problems in elementary school-age children and could be used for various educational purposes, including research projects and intervention programs. Hence, the MBC can be considered as an effective method for assisting teachers and physical education instructors in referring students for a more comprehensive assessment, as well as for collecting data from children and youth as part of the clinical investigation process. This is because the behavioral patterns of children with ADHD and ASD can be more clearly expressed in social contact environments with different stimuli than in the classroom environment where stimuli are closely regulated, and students are expected to obey more strict rules of behaviors (Efstratopoulou et al. 2012b).

1.8. Previous Cross-Cultural Adaptation of MBC in Other Countries

Even though in the relevant literature there are contradictive views regarding the procedures that are necessary for a cross-cultural adaptation of assessment scales, scholars agree that the process must go beyond simple translation, since translation alone does not ensure the instrument's reliability and construct validity (International Test Commission 2017). For example, Borsa et al. (2012) have suggested a six-step process of cross-cultural adaptation. These steps include (1) translation of the instrument from the source language into the target language; (2) synthesis of translated versions; (3) synthesis evaluation by expert judges; (4) evaluation of the instrument by the target groups; (5) back-translation; and (6) a pilot study. In addition, the authors also emphasized the importance of assessing the factorial structure of the instrument to confirm its stability concerning the original document. Alternatively, Mondrzak et al. (2016) used the guidelines proposed by the 'Task Force for Translation and Cultural Adaptation of the International Society for Pharmacoeconomics and Outcomes Research'. This 10-step process can be described as follows: preparation, forward translation, reconciliation of different translations into a single version, back-translation, back-translation review, harmonization, cognitive debriefing, review of cognitive debriefing results and finalization, proofreading, and final report. Wild et al. (2005) have also proposed a similar process.

Previous studies on the evaluation of the psychometric properties of MBC for children have revealed that the MBC is a content-homogeneous instrument, with high temporal stability and high interrater agreement that can provide useful and reliable ratings on behavioral and emotional problems in children, especially when used by PE teachers in school settings (Efstratopoulou et al. 2012a, 2012b, 2013, 2015) (Table A1).

Up to date, the Motor Behavior Checklist (MBC; Efstratopoulou et al. 2012a) has already been translated into six languages (Greek, Polish, Urdu, Czech, Chinese, Brazilian/Portuguese) and has been used in several studies (Paiano et al. 2019; Wood and Efstratopoulou 2020; Efstratopoulou et al. 2017).

2. The Present Study

Recognizing the need for incorporating such an innovative screening tool (MBC) in the area of early diagnosis of neurodevelopmental, behavioral, and developmental disorders in the Arabic population, the researchers in the present study followed the ten-step procedure suggested by Wild et al. (2005) and Mondrzak et al. (2016), to translate and culturally adapt the MBC to the Arabic language and UAE culture. The main aim of the study is to provide a culturally validated version of MBC in the target language, which will assist clinicians and multidisciplinary groups working in the field of Special Education in carrying out more accurate, valid, and complete screening procedures and diagnoses of students, mainly those who might be 'at risk' of NDD, in the United Arab Emirates.

3. Method

Research that crosses linguistic and cultural boundaries necessarily requires direct attention to the use of language and cultural factors where verbal expressions, comprehension, or both are involved at any level in the systematic collection of data expected to exhibit comparable reliability and validity across the linguistic and cultural boundaries.

Idiomatic criteria were used to assess the equivalence with the new version. The analysis used and the estimation of contextual and cultural sensitivity was based on culture-loaded phraseological expressions that were used to exhibit strong contextuality. The structural equivalence of the list (and each sub-factor) was demonstrated between native speakers of the target language. It was important to test the measurement equivalence between native speakers of the source language and the target language. If the latter is not supported, the underlying reasons and cultural factors causing the lack of equivalence were further discussed.

Characteristics of the Sample

The Physical Education (PE) teachers who participated in this study were native (Arabic) speakers with a specialization in Adapted Physical Education (APA) for children with disabilities. The criteria used for the selection of the professionals who participated in the Focus Group were: (a) To be native Arabic speakers, (b) To have at least four years of teaching experience in public primary schools in UAE, (c) To have a specialization in APA and knowledge of the characteristics of children with disabilities and of behavioral management interventions. For our sample, the professionals were all native speakers, had MN = 5.6 (SD = 1.2) years of teaching experience in Primary Schools in UAE and they all had a Degree in Physical Education with a concentration in Adapted Physical Activity (from the United Arab Emirates University). In addition, all of them had practical experience working in public (both mainstream and special) schools in the UAE, with children with disabilities.

The translation and cross-cultural adaptation processes used in this study, as previously mentioned, followed the major guidelines suggested in the works of Wild et al. (2005) and Mondrzak et al. (2016). The process included ten steps, as illustrated in Figure 1 and Table 1. The Ethics committee of United Arab Emirates University (UAEU, Protocol No ERS_2021_7335) approved the project.

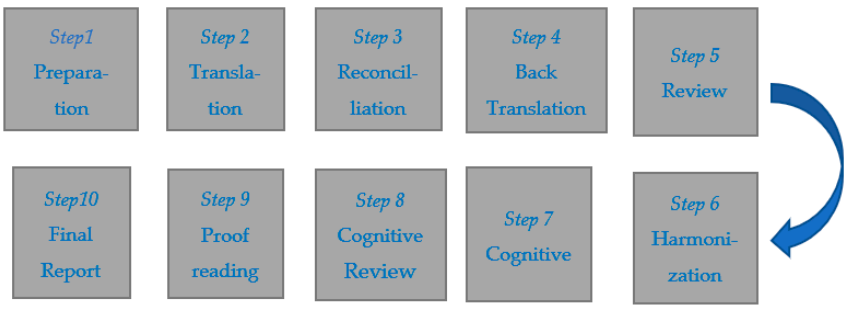

Figure 1. Steps involved in the translation and cross-cultural adaptation process of the MBC.

Table 1. Detailed description of the translation and cross-cultural adaptation process of the MBC.

Steps	Actions
1	**Preparation:** The project manager (who is also the MBC developer), recruited three key persons from the target country to work on the project. Permission for translation was acquired. The instrument developer provided information and clarifications on the conceptual basis of the instrument items for use by the translators. A key person from the target country worked closely with the project manager for the preparation of the translation process.
2	**Forward Translation:** Three professionals specializing in Special Education—native speakers of the target language (Arabic), with proficiency in English, independently translated the instrument into Arabic. The project manager worked closely with the translators to provide background information about the conceptual basis of the instrument and/or the particular wordings to be used in the items.
3	**Reconciliation of the forward translations into a single forward translation:** After completion of the individual translations, the translators compared the different versions of each item to reach a consensus, aiming at the most appropriate cultural adjustment and resolving possible discrepancies.
4	**Back translation of the reconciled translation into the source language:** The instrument was back-translated into the source language by an English teacher with proficiency in both the English and the Arabic language (i.e., an English native speaker living in the UAE).
5	**Back translation review:** To ensure the conceptual equivalence of the translation, the project manager and the key in-country person reviewed the back translation to identify any discrepancies or problematic items and to decide upon the best linguistic and semantic match between the wordings in the two versions. A professor of Arabic Language worked on the Arabic version during the back translation review procedure.
6	**Harmonization between all translated versions and with the source version:** This is an additional quality-control step to further ensure that all linguistic or conceptual discrepancies are resolved. Thus, based on the comments of the back-translator, the authors designed a new synthesized version (aggregation of the global data set).
7	**Cognitive debriefing:** to assess the level of comprehensibility and cognitive equivalence of the final translated version, four native speakers of the target language, specializing in the target area (i.e., Physical Education (PE) teachers *, working with school-aged children), evaluated the translated instrument. Additional issues causing confusion were resolved during this phase.
8	**Cognitive debriefing review:** The project manager reviewed the results from the cognitive debriefing and identified translation modifications necessary for improvement. Partially clear items were analyzed by a focus group composed of three Physical Education (PE) teachers *. Following agreement on changes between the project manager and the key in-country persons, the translation process was finalized.
9	**Proofreading:** The finalized translation was proofread to check and correct any remaining spelling, grammatical, or other errors. The final suggestions and corrections were sent for approval by the author of the original version.
10	**Final Report:** The project's Principal Investigator developed the final report, which included a full description of the methodology used, along with an item-by-item representation of all translation decisions undertaken throughout the process. Finally, the author's comments were analyzed, and the final version of the instrument was produced.

* The PE teachers participated in this study were native (Arabic) speakers with specialization in Adapted Physical Education (APA) for children with disabilities.

4. Results

The comparison between the back-translated version and the original version (step 4) revealed that 44 items were translated at a satisfactory level (78%). Fifteen of the items (22%), which were considered divergent, were forwarded to the back-translator, along with the original version, for further consideration. The back-translator characterized nine of these items as 'semantically identical'—-and, therefore, acceptable—-, but commended on the other six items, setting up step 5 (Table 2). Based on the back translators' comments, a further synthesized version of the instrument was developed (step 6). The next step (cognitive debriefing) revealed that 84.7% of the items were comprehensive and cognitively equivalent (i.e., 71% had a mean score of 2.87, while 15 items 18.6% had a mean score of 2.35). Nine items were considered partially clear (i.e., three items with a mean score of 2.5, three items with a mean score of 2.05, and two items with a mean score of 2.01). The overall results of the back-translated version analysis, along with the modifications proposed by the focus group (step 8) are illustrated in Table 2.

Table 2. Comparison between the original version of the MBC checklist and the translated and back-translated versions with semantic adaptation.

Original Version	Arabic Translation	Back Translation to English	Back- Translator's Notes
Rules Breaking	**كسر القواعد**	**Breaking the rules**	
15. Is aggressive towards leadership figures	عدواني تجاه الشخصيات القيادية (المدراء أو المعلمين)	He/ she is aggressive towards leading figures (for example, school principals, teachers)	Adding examples to clarify meaning .Minor adaptation
27. Blames others for his/her mistakes	يلوم الآخرين على الأخطاء التي ارتكبها	He blames others for his own mistakes	Added to clarify the meaning Minor adaptation
II. Hyperactivity Impulsivity	**فرط الحركة / الاندفاعية**	**Hyperactivity / Impulsivity**	
2. Has difficulty waiting his/her turn to perform	لديه صعوبة في انتظار دوره عندما يشارك المجموعة	He/she has a hard time waiting for his /her turn when participate in group activities.	Added to clarify the meaning of the word 'perform' - Minor adaptation
16. Interrupts others (e.g., butts into conversation)	يقاطع الآخرين أثناء التحدث	Interrupts others while speaking	- Minor adaptation
22. Interrupts others (e.g., butts into games)	يقاطع الآخرين أثناء اللعب	Interrupts others while playing	- Minor adaptation
28. Doesn't care for equipment	لا يهتم بـ أو يحطم الأدوات المستخدمة في الأنشطة داخل الفصل أو في حصص الرياضة	Doesn't care for or destroy tools used for in-class activities or in sports	- Major adaptation based on teacher's comments
33. May shift from one uncompleted activity to another	ينتقل من نشاط لأخر دون أن يكمل أي منهما	He/she moves from one activity to another without completing either of them	- Minor adaptation
42. Appears to be 'driven by a motor'	يتحرك باستمرار كما لو كان لديه محرك – مفرط النشاط	He /she is constantly moving as if he/ she 'has an engine' is hyperactive	Major adaptation based on teacher's comments
Lack in Attention	**نقص في الانتباه**	**Lack of attention**	
43. Makes careless mistakes in activities	يرتكب أخطاء لعدم الانتباه أو اللامبالاة أثناء تنفيذ الأنشطة	Makes mistakes due to inattention or indifference while carrying out activities	Added to clarify the meaning Minor adaptation
55. Avoids or has strong dislike for activities that require organizational demands	يتجنب أو لديه كراهية شديدة للأنشطة التي تتطلب التنظيم والانتباه المستمر	Avoids or has a extreme hatred for activities that require regulation and sustain attention	Minor adaptation
VI. Lack in Social Interaction	**نقص في التفاعل الاجتماعي**	**Lack of social interaction**	
6. Displays impairment in gestures which regulate social interaction	يُظهر عجز في استخدام الإيماءات الجسدية (لغة الجسد) التي تنظم التفاعل الاجتماعي	Demonstrates a deficit in the use of physical gestures (body language) that regulate social interaction	Added to clarify the meaning Minor adaptation
19. Doesn't show objects he/she finds interesting	لا يعرض الأشياء التي تثير اهتمامه عندما يُطلب منه ذلك	Does not show things that interest him when asked to do so	Added to clarify the meaning Minor adaptation
25. Doesn't bring objects he/she finds interesting	لا يجلب أو يحضر الأشياء التي تثير اهتمامه عندما يطلب من ذلك	He does not bring things that interest him/ her when asked to do so	Added to clarify the meaning Minor adaptation
48. Doesn't want physical contact	لا يريد الاتصال الجسدي أو التقارب الجسمي	He does not want physical contact or physical closeness	Added to clarify the meaning Minor adaptation
VII. Lack of self-regulation	**سابعا. نقص التنظيم الذاتي**	**VII. Lack of self-regulation**	
4. Has difficulties organizing activities	لديه صعوبات في تنظيم الأنشطة التي يشارك فيها	Has difficulties organizing activities in which he/she participates	Added to clarify the meaning Minor adaptation

More specifically, most of the items were considered clear by the focus group and, therefore, required no alterations. Items 28 and 42—-included in the hyperactivity factor—-were characterized as 'not clear enough' and, therefore, the focus group and the back translator expert proposed major adaptations. These items were modified based on the recommendations from the focus group, to make sure that professionals in school settings in the UAE can observe and record these specific behaviors in their students during class activities and/or during free play situations. The rest of the items either remained the same as in the preliminary translated version or needed minor modifications, which, in most cases, was either the addition/replacement of a verb or the addition of an example to clarify the meaning and make sure that the rater is satisfactorily assessing students' externalizing and internalizing behaviors in school settings.

5. Discussion

The present study described the processes followed during the translation and cultural adaptation of the MBC into the Arabic language. The new Arabic version of the MBC was developed to provide class teachers and physical education teachers in the UAE, with a practical, easy to administer, useful, and valid assessment tool to observe and record the motor behavior of their 6 to 12-years-old students. As aforementioned, according to Gardon (2012), teachers are considered as the most representative and valid source of information regarding their students' manifestation of behavioral problems. Supported by valid and reliable assessment instruments, professionals in education can provide detailed reports on children's deviant behavior observed in natural conditions of interaction and competition, which are rarely considered in evaluation protocols.

Educational professionals can use the MBC for screening and assessment purposes of children with emotional and/or behavioral disorders (EBP) and/or developmental disorders, during class activities physical activity classes and free-play situations. As research suggests, EBP are often highly prevalent among children and adolescents with neurodevelopmental disorders (e.g., autism spectrum disorder (ASD), Attention-Deficit/Hyperactivity Disorder (ADHD), Specific Learning Disorder (SLD), etc.) (Eisenberg et al. 2000). More specifically, it has been found that the specific behavioral patterns of children with ADHD and ASD can be more clearly expressed in social contact environments, such as free-play or sports, where stimuli are more closely regulated and students are expected to obey more strict rules of behaviors, rather than in the typical classroom environment (Efstratopoulou et al. 2012b). In addition, as Efstratopoulou et al. (2019) have pointed out, a detailed assessment, performed by both professionals and teachers using multiple instruments, can safeguard a proper identification of behavioral changes and/or developmental delays.

Further discussing the importance of cultural adaptations made in this study concerning the Arabic version of MBC for use by PE teachers and professionals in school settings, we need to point out that all adaptations made, were based on recommendations from professionals in UAE who are working with students in primary school settings. The professionals, who participated at the focus group, indicated that there was a specific need for some of the items to reflect the cultural environment in schools, taking into consideration that not all students' deviant behaviors are easy to be observed in a structured class environment in UAE schools. Furthermore, the significance of cross-cultural adaptation of the MBC for use in different countries has also been demonstrated in previous studies (e.g., Paiano et al. 2019; Efstratopoulou et al. 2012a). Overall results are in accordance with other studies, which have also revealed that, based on the synthesized version, the translation and back-translation processes are adequate methods for translating and culturally adapting instruments, without major distortions (Mondrzak et al. 2016; Wild et al. 2005; Borsa et al. 2012; Gjersing et al. 2010; Mattos et al. 2006).

Practical Implications and Recommendations for Future Research

Considering the above, the newly translated and culturally adapted version of the MBC, as described in this article, will comprise a useful tool for the assessment of children

at risk of EBP and/or NDD, by physical education teachers and/or classroom teachers in the UAE. Furthermore, results revealed a satisfactory level of agreement between the original and back-translated versions, with 78% of exact equivalence between the translated items and 12% of terms requiring minor adjustments. In addition, clarity assessment using reports from teachers revealed an 84% agreement with the draft version of the MBC. The synthesized version of the instrument required modifications to ensure semantic and cultural adequacy in relation to the original version. Minor adaptations based on recommendations from professionals in UAE participating at the focus group, indicated that there was a need for some items to reflect the cultural environment in schools taking into consideration that not all students' deviant behaviors are easy to be observed in a structured class environment in UAE schools. For example, for the items 19 and 25 (behaviors that are connected with lack of social skills), the focus group suggested the addition of the phrase: *when asked to do so* (which is missing from the original English version of the MBC checklist) but was proposed by professionals/educators in UAE) with the explanation that these social behaviors are not observable in UAE schools settings unless there is a specific request from the educators (class teachers/PE teachers).

It was also mentioned that due to schools' rules and maybe cultural components, children in schools do not have the flexibility to express in different ways in a structure class environment unless they were asked or motivated to do so. In general, the Arabic version of the instrument showed adequate indicators of semantic equivalence following the steps of initial translation, back-translation, and clarity assessment by professionals and by the focus group.

Future research is needed to assess the psychometric properties (mainly the validity and reliability) of the new Arabic version of the MBC using a large sample of primary school-aged children from UAE schools rated by their teachers and physical educators in school settings. More specifically, future studies need to collect data from typical and clinical samples of children from UAE primary schools to assess the psychometric properties of the checklist and to ensure that the Arabic version of MBC is a new valid and reliable assessment instrument to support teachers, physical educators, and special educators in their role in UAE schools. In addition, evidence of validity should be demonstrated through multiple informants. A study currently in progress, by our group, has started to assess the psychometric properties of the translated version of the checklist.

Hence, the new instrument fills a gap in the evaluation process of students in sports and free-play situations. Moreover, it can help schoolteachers and physical educators to better understand and effectively deal with their students' behavioral profiles, especially those with behavior problems compatible with NDD. Consequently, we highlight the importance of the cross-cultural adaptation of instruments for use in different countries, as demonstrated by previous works and performed in our study.

6. Conclusions

The cross-cultural adaptation and translation processes used in this article allowed the formulation of an Arabic version of the Motor Behavior Checklist for children (MBC; Efstratopoulou et al. 2012b) that will enable physical education teachers to evaluate their students' behavioral aspects in sports and free-play situations. The Arabic version of the MBC was produced following rigorous translation and cross-cultural adaptation procedures. Results from the pilot use of the MBC in the UAE indicated that there is a satisfactory level of agreement between the original and the back-translated versions. The final translated version's overall content was found sufficiently compatible with the original version of the instrument and was proven to ensure a more complete and comprehensive evaluation process. Teachers, PE teachers, and educational professionals will be able to provide valid reports on students' behaviors, observed in natural school settings of interaction and competition, such as sports and free-play situations, which are rarely considered in evaluation protocols. Moreover, it can help schoolteachers to better understand and effectively deal with their students' behavioral profiles, especially those with behavioral problems

compatible with NDD. Future research is required to further assess the psychometric properties of the new Arabic version of the MBC using a larger sample of primary school-aged children from UAE schools rated by their teachers and physical educators in school settings. Therefore, we highlight the importance of the cross-cultural adaptation of instruments for use in different countries, as demonstrated by previous works and performed in our study.

Author Contributions: Conceptualization, M.E. and H.E.; methodology, M.E. and H.E.; software, M.E., H.E. and A.A.E.; validation, M.E., H.E. and E.B.; formal analysis, M.E. and H.E.; investigation, M.E., H.E., A.A.E. and E.B.; resources, M.E., H.E. and A.A.E.; data curation, M.E. and H.E.; writing—original draft preparation, M.E., H.E. and E.B.; writing—review and editing M.E., H.E. and E.B.; visualization, M.E., H.E. and E.B.; supervision, M.E., H.E. and E.B.; project administration, M.E., H.E. and E.B. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported in part by the Research Office in UAEU (Number of grant G00003523/2021). The authors wish to thank the UAEU for its financial support.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of United Arab Emirates University (UAEU) (protocol code: ERS_2021_7335 and date of approval: 20/6/2021).

Informed Consent Statement: Informed consent was obtained from all participants involved in the study. Written informed consent has been obtained from the participants to publish this paper.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy issues.

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

Appendix A

Table A1. The Arabic Version of MBC after translation and cultural Adaptations.

		إطلاقا	أحيانا	في أغلب الأحيان	في أكثر أغلب الأحيان	تقريبا دائما
1	لا يلتزم بالقواعد خاصة أثناء اللعب	0	1	2	3	4
2	لديه صعوبة في انتظار دوره عند القيام بالأنشطة	0	1	2	3	4
3	غير مهتم أو مهمل(مثلا في أداء الأنشطة)	0	1	2	3	4
4	يَظهر عليه التعب بعد مجهود بسيط.	0	1	2	3	4
5	يعرض حركات جسمية نمطية، والتي تشمل اليدين (مثل التصفيق ونقر الأصابع)	0	1	2	3	4
6	يُظهر عجز في استخدام الإيماءات التي تنظم التفاعل الاجتماعي	0	1	2	3	4
7	يُظْهر الانشغال المستمر بأجزاء الأشياء	0	1	2	3	4
8	يظهر العصيان لمعلمه	0	1	2	3	4
9	يظهر فرط الحركة أثناء الدرس	0	1	2	3	4
10	لديه صعوبة في التركيز	0	1	2	3	4
11	يشعر بالدوار، أو عدم الاستقرار، أو الإغماء	0	1	2	3	4
12	يعرض أنماط متكررة من الأنشطة	0	1	2	3	4
13	يتجنب الأنشطة الاجتماعية المناسبة لعمره	0	1	2	3	4
14	لا يظهر أي اهتمام للدرس	0	1	2	3	4
15	عدواني تجاه الشخصيات القيادية (المعلمين أو الزملاء)	0	1	2	3	4
16	يقاطع الآخرين (على سبيل المثال أثناء التحدث)	0	1	2	3	4
17	لديه صعوبة في الحفاظ على الانتباه في المهام	0	1	2	3	4
18	يُظهر قلة الطاقة	0	1	2	3	4
19	لا يَعْرِض الأشياء التي يجدها ممتعة	0	1	2	3	4
20	يظهر عجز ملحوظ في استخدام السلوكيات غير اللفظية مثل الاتصال البصري	0	1	2	3	4

Table A1. *Cont.*

		إطلاقا	أحيانا	في أغلب الأحيان	في أكثر أغلب الأحيان	تقريبا دائما
21	سلبي تجاه زملائه في الفصل	0	1	2	3	4
22	يقاطع الآخرين (على سبيل المثال أثناء اللعب)	0	1	2	3	4
23	يبدو كما لو أنه لم يسمع ما قيل للتو	0	1	2	3	4
24	يُظهر مستوى منخفض من النشاط	0	1	2	3	4
25	لا يُحْضِرُ الأشياء التي يجدها ممتعة	0	1	2	3	4
26	يظهر عجز في تعبيرات الوجه	0	1	2	3	4
27	يلوم الآخرين على أخطائه	0	1	2	3	4
28	لا يهتم بالأدوات المستخدمة في اللعب	0	1	2	3	4
29	يتجنب أو لديه كراهية شديدة للأنشطة التي تتطلب تركيزًا شديدا	0	1	2	3	4
30	ينسحب من الاتصال بالآخرين	0	1	2	3	4
31	لا يدرك أن خوفه زائد عن الحد	0	1	2	3	4
32	يلعب بقسوة أثناء الألعاب الجماعية	0	1	2	3	4
33	قد ينتقل من نشاط غير مكتمل إلى آخر	0	1	2	3	4
34	يظهر صعوبة التركيز في بداية الدرس	0	1	2	3	4
35	يُظْهر نقص في التواصل مع زملائه في الفصل	0	1	2	3	4
36	يخاف من الوقوف في الطابور	0	1	2	3	4
37	يظهر نزعة التنمر تجاه زملائه في الفصل	0	1	2	3	4
38	ينخرط في أنشطة خطرة دون مراعاة العواقب المحتملة	0	1	2	3	4
39	لديه صعوبات في تنظيم المهام	0	1	2	3	4
40	ينعزل من قبل زملائه في الفصل	0	1	2	3	4
41	يُظهر القلق الذي قد يتم التعبير عنه بالبكاء أو نوبات الغضب أو التجميد أو التشبث	0	1	2	3	4
42	يبدو أن هناك محرك يقود الطالب	0	1	2	3	4
43	يرتكب أخطاء لعدم الانتباه أو اللامبالاة أثناء تنفيذ الأنشطة	0	1	2	3	4
44	يحاول البقاء على مقربة من الكبار المألوفين له	0	1	2	3	4
45	يظهر صعوبة اتخاذ القرارات	0	1	2	3	4
46	يجد صعوبة في اللعب أو الانخراط بهدوء في الأنشطة الترفيهية	0	1	2	3	4
47	لا يهتم بالتفاصيل	0	1	2	3	4
48	لا يريد الاتصال الجسدي أو التقارب الجسمي	0	1	2	3	4
49	يجد صعوبة في السيطرة على القلق	0	1	2	3	4
50	ينزعج عندما يخسر	0	1	2	3	4
51	لا يشارك بنشاط في اللعب الاجتماعي البسيط	0	1	2	3	4
52	ينزعج عندما يفشل في إنجاز مهمة ما	0	1	2	3	4
53	يبالغ في تقدير قدراته	0	1	2	3	4
54	لديه صعوبات في تنظيم الأنشطةالتي يقوم بها	0	1	2	3	4
55	يتجنب أو لديه كراهية شديدة للأنشطة التي تتطلب متطلبات تنظيمية أو متعلقة بالترتيب والتنظيم	0	1	2	3	4
56	يظهر السلوك المتهور	0	1	2	3	4
57	يلمس الأشياء التي ليس من المفترض أن يلمسها	0	1	2	3	4
58	يظهر نقص في اللعب التخيلي المتنوع	0	1	2	3	4
59	يفقد أعصابه	0	1	2	3	4

References

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Arlington: American Psychiatric Association.

Borsa, Juliane Callegaro, Bruno Figueiredo Damásio, and Denise Ruschel Bandeira. 2012. Adaptação e validação de instrumentospsicológicos entre culturas: Algumasconsiderações. *Paideia* 22: 423–32. [CrossRef]

Efstratopoulou, Maria, Rianne Janssen, and Johan Simons. 2012a. Agreement among physical educators, teachers and parents on children's behaviors: A multitrait-multimethod design approach. *Research in Developmental Disabilities* 33: 1343–51. [CrossRef] [PubMed]

Efstratopoulou, Maria, Rianne Janssen, and Johan Simons. 2012b. Differentiating children with attention-deficit/hyperactivity disorder, conduct disorder, learning disabilities and autistic spectrum disorders by means of their motor behavior characteristics. *Research in Developmental Disabilities* 33: 196–204. [CrossRef] [PubMed]

Efstratopoulou, Maria, Johan Simons, and Rianne Janssen. 2013. Concordance among physical educators', teachers', and parents' perceptions of attention problems in children. *Journal of Attention Disorders* 17: 437–43. [CrossRef] [PubMed]

Efstratopoulou, Maria, Rianne Janssen, and Johan Simons. 2015. Assessing children at risk: Psychometric properties of the motor behavior checklist. *Journal of Attention Disorders* 19: 1054–63. [CrossRef] [PubMed]

Efstratopoulou, Maria, Thomas Dunn, Agnieszka Augustyniak, and Joanna Andrzejewska. 2017. Assessing externalizing and internalizing behavior in children: Use of the Motor Behavior Checklist in a typical school-aged Polish sample. *European Journal of Special Education Research* 2: 38–46.

Efstratopoulou, Maria, Rianne Janssen, and Johan Simons. 2019. Children's deviant behavior in primary education: Comparing physical educator's implicit theory with diagnostic criteria. *Journal of Attention Disorders* 23: 246–56. [CrossRef] [PubMed]

Eisenberg, Nancy, Richard A. Fabes, Ivanna K. Guthrie, and Mark Reiser. 2000. Dispositional emotionality and regulation: Their role in predicting quality of social functioning. *Journal of Personality and Social Psychology* 78: 136–57. [CrossRef] [PubMed]

Gardon, Lyn. 2012. The school behaviors rating scale: A measure to assess behavioral strengths and needs and inform supportive programming. In *Transforming Troubled Lives: Strategies and Interventions for Children with Social, Emotional and Behavioral Difficulties*. Bradford: Emerald Group Publishing Limited, vol. 2, pp. 75–92. [CrossRef]

Gjersing, Linn, John R. Caplehorn, and Thomas Clausen. 2010. Cross-cultural adaptation of research instruments: Language, setting, time and statistical considerations. *BMC Medical Research Methodology* 10: 13. [CrossRef] [PubMed]

Haack, Lauren Marie Haack, Kelsey Gonring, Michael Harris, Alyson Gerdes, and Linda Pfiffner. 2019. Assessing impairment in childhood ADHD: Validation of the parent and teacher ADHD-fx rating scale in a dual-site clinical sample. *Journal of Attention Disorders* 23: 541–52. [CrossRef] [PubMed]

International Test Commission. 2017. The ITC Guidelines for Translating and Adapting Tests (Second Edition). Available online: https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf (accessed on 23 September 2021).

Jensen, Peter S., Stephen P. Hinshaw, Helena C. Kraemer, Nilantha Lenora, Jeffrey H. Newcorn, Howard B. Abikoff, John S. March, L. Eugene Arnold, Dennis P. Cantwell, C. Keith Conners, and et al. 2001. ADHD comorbidity findings from the MTA study: Comparing comorbid subgroups. *Journal of the American Academy of Child and Adolescent Psychiatry* 40: 147–58. [CrossRef] [PubMed]

Kashani, Javad H., Wesley D. Allan, Niels C. Beck, Yolanda Bledsoe, and John C. Reid. 1997. Dysthymic disorder in clinically referred preschoolchildren. *Journal of the American Academy of Child and Adolescents Psychiatry* 36: 1426–33. [CrossRef] [PubMed]

Mattos, Paulo, Daniel Segenreich, Eloísa Saboya, Mário Louzã, Gabriela Dias, and Marcos Romano. 2006. Adaptação transcultural para o português da escala Adult Self-Report Scale para avaliação do transtorno de déficit de atenção/hiperatividade (TDAH) emadultos. *Revista de PsiquiatriaClinica* 33: 188–94. [CrossRef]

Mol Lous, Annemieke, Cees A.M. de Wit, Eric E. J. de Bruyn, and J. Marianne Riksen-Walraven. 2002. Depression markers in young children's play: A comparison between depressed and non-depressed 3 to 6 years old in various play situations. *Journal of Child Psychology and Psychiatry* 43: 1029–38. [CrossRef] [PubMed]

Mondrzak, Rafael, Camila Reinert, Andreia Sandri, Lucas Spanemberg, Eduardo L. Nogueira, Mirella Bertoluci, Claudio Laks Eizirik, and Nina Rosa Furtado. 2016. Translation and cross-cultural adaptation of the rating scale for countertransference (rsct) to American English. *Trends in Psychiatry and Psychotherapy* 38: 221–26. [CrossRef] [PubMed]

Paiano, Rone, Maria Christina Triguero Veloz Teixeira, Carla Nunes Cantiere, Maria A. Efstratopoulou, and Luis Renato Rodrigues Carreiro. 2019. Translation and cross-cultural adaptation of the motor behavior checklist (MBC) into Brazilian Portuguese. *Trends of Psychiatry and Psychotherapy* 41: 167–75. [CrossRef] [PubMed]

Wild, Diane, Alyson Grove, Mona Martin, Sonya Eremenco, Sandra McElroy, Aneesa Verjee-Lorenz, Pennifer Erikson, and ISPOR Task Force for Translation and Cultural Adaption. 2005. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (pro) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value in Health* 8: 94–104. [CrossRef] [PubMed]

Wood, Sue, and Maria Efstratopoulou. 2020. Assessment in Special Education: Improving professionals' skills using video footage. In *Interdisciplinary Reflections on Sociocultural issues in Education*. Edited by Alexandros Argiriadis. Cambridge: Cambridge Scholar Publishing.

Journal of
Intelligence

MDPI

Article

Parental Stress and Children's Self-Regulation Problems in Families with Children with Autism Spectrum Disorder (ASD)

Maria Efstratopoulou [1], Maria Sofologi [2,3], Sofia Giannoglou [4] and Eleni Bonti [4,5,*]

[1] Department of Special Education (CEDU), United Arab Emirates University (UAEU), Al Ain P.O. Box 15551, United Arab Emirates; maria.efstratopoulou@uaeu.ac.ae
[2] Laboratory of Psychology, Department of Early Childhood Education, School of Education, University of Ioannina, 45100 Ioannina, Greece; m.sofologi@uoi.gr
[3] Institute of Humanities and Social Sciences, University Research Centre of Ioannina (URCI), 45110 Ioannina, Greece
[4] First Psychiatric Clinic, School of Medicine, Faculty of Health Sciences, Aristotle University of Thessaloni-ki, "Papageorgiou" General Hospital, Ring Road Thessaloniki, N. Efkarpia, 54603 Thessaloniki, Greece; sophiegianno@gmail.com
[5] Department of Education, School of Education, University of Nicosia, Nicosia 2417, Cyprus
* Correspondence: bonti@auth.gr

Abstract: *Background*: Increased parental stress is strongly related to the severity of autism spectrum disorder (ASD) symptomatology. Parents' coping strategies and social support issues add to the complexity of this relationship. *Aim*: The present study investigated the relationship between self-regulation skills and parenting stress in parents of nonverbal children with ASD. *Methods and procedure*: The Parenting Stress Index–Short Form (PSI-SF) was administered to 75 families, and self-regulation scores on a Motor Behavior Checklist for children (MBC) were recorded by students' class teachers (level of functioning-behavioral problems). In addition, interviews were conducted with a focus group of six parents (four mothers and two fathers) to explore the underline factors of parental stressin-depth. *Results*: Correlation analyses revealed that parenting stress was positively correlated with elevated scores on MBC children's self-regulation subscale. On the other hand, parenting stress was negatively correlated with the level of social functional support reported. Qualitative data were analyzed using transcripts, revealing additional stressors for families and parents, and resulting in recommendations to overcome these factors. *Conclusions and implications*: Aiming at developing strategies to improve self-regulation skills in nonverbal children with ASD may be particularly important in reducing parental stress for families having nonverbal children with autism and other developmental disabilities. Parents' stressors and suggestions during interviews are also discussed.

Keywords: parental stress; self-regulation; social support; coping strategies; ASD; behavioral difficulties; non-verbal children

Citation: Efstratopoulou, Maria, Maria Sofologi, Sofia Giannoglou, and Eleni Bonti. 2022. Parental Stress and Children's Self-Regulation Problems in Families with Children with Autism Spectrum Disorder (ASD). *Journal of Intelligence* 10: 4. https://doi.org/10.3390/jintelligence10010004

Received: 16 November 2021
Accepted: 13 January 2022
Published: 17 January 2022

1. Introduction

Family schema is a dynamic living cell. Parents as caretakers play a vital role inmentoring and guiding their children's learning journey from early childhood into adulthood. Parental stress is an anxiety closely aligned with the significant role of being a parent. However, when it comes to parents of children with autism spectrum disorder (ASD), this role often starts as soon as their child is diagnosed (Foster et al. 2012). More specifically, parents of children with ASD very often face a plethora of difficulties regarding the educational opportunities focused on developing their skills and achieving a high quality of life (Meadan et al. 2010; Tincani et al. 2014). In this vein, parents are faced with a range of emotional pressures as they attempt to learn about ASD and what this means for their child. Compared with parents of children with intellectual or other developmental disabilities, studies reveal that parents of children with ASD experience more psychological distress,

including depression, anxiety, and components of stress, such as decreased family cohesion and increased somatic symptomatology (Martins et al. 2015).

1.1. Autism Diagnosis and Parenting Stress

An autism diagnosis not only changes the life of the child diagnosed but also the quality of emotional harmony of all family members. Parents describe experiencing initial feelings of surprise, sadness, shock, and rejection following their child's diagnosis (Martins et al. 2015). It is an axiom to realize that the family, as a holistic schema, is facing and must cope with the challenges of communicating with and assisting the social interactions of their child with ASD. At this point, it is essential to underline that parents have to manage high levels of anxiety owing to perplexing therapeutical programs, such as home-implemented treatments and juggling job responsibilities, as well as family commitments (Hayes and Watson 2013). Thus, the diagnosis of a child with autism spectrum disorder (ASD) can be a time of tremendous uncertainty, agony, and a depressive journey for parents and families. Parents must handle a maze of information and bureaucratic processes as they attempt to find the appropriate therapeutic intervention program to strongly support their child (Tomeny 2017).

There is a great variety of therapeutic approaches, which focus on accommodating or remediating different challenges related to autism. The therapeutic approach is effective only when it is individualized depending on the chronological age and developmental level; therapies focus on all difficulties, the interactions between them, and the promotion of growth and adaptation in the long run (Poppi et al. 2019). In Greece, around the time that parents start to come to terms with their child's condition, they have to decide what kind of treatment plan they are going to follow. Children with autism are considered to be very different from each other, and the clinical presentation of their symptoms varies along with the outcomes following an intervention (Ben-Itzchak et al. 2014). There are many different treatments provided to children with autism, such as the treatment and education of autistic and related communication-handicapped children (TEACCH), cognitive behavioral therapy, applied behavior Analysis (ABA), learning experiences: an alternative program for preschoolers and parents (LEAP),picture exchange communication systems (PECS),speech and language therapy, occupational therapy, etc.(Poppi et al. 2019; Ben-Itzchak et al. 2014; CDC 2021).

Several significant factors are closely aligned with anxiety or agony in parents of children with ASD (e.g., child characteristics, lack of emotional and social support) (Lee et al. 2009). Although research reveals an impact on family members of individuals with ASD, Hastings et al. (2005a) emphasize that not all family members experience similar effects as a result of having an individual with ASD in the family. For example, Hastings (2003) found that mothers of children with ASD reported more anxiety and negative outcomes than fathers in the same family. In addition, researchers have found positive outcomes (e.g., limited conflicts within the relationship, high self-esteem and self-concept) for some typically developing siblings of individuals with ASD (Sharabi and Marom-Golan 2018), whereas some parents described the experience of having a child with ASD as being positive (Hastings et al. 2005a).

Under the scope of evaluating parental levels of anxiety, we must underline the fact that, although the research community has focused on subgroups referred to as "high-functioning autism" (Baio et al. 2018), the influence of the child's IQ (intellectual quotient) level on family functioning has hardly been studied. Around 70% of parents of children with Asperger syndrome score at or above the 90th percentile of the normal parental stress scores (Matson and Kozlowski 2011). It appears that a high level of cognitive development in children with ASD does not, in itself, influence and reduce the stress produced by raising a child with ASD. There is a research consensus emphasizing the fact that parents of children with high-functioning autism report significantly higher levels of long-term agony and lower levels of adaptive coping strategies and resources when compared with parents of children with typical development (Hayes and Watson 2013). Furthermore, concerning

the lifelong nature of ASD, these challenges are often longstanding and extend into the sons' and daughters' adolescense and adulthood. Moreover, parents confirm that much of their stress and emotional exhaustion is caused by the continued necessity of having to fight for services, cope with complicated policies or negative societal attitudes, and constantly having to communicate and build relationships with education and health professionals (Yorke et al. 2018).

Undoubtedly, an increased tendency to experience negative emotions and a decreased ability to regulate emotional responses appear common to many childhood psychiatric and developmental disorders—although, not surprisingly, the links between other temperamental traits and psychopathology vary by disorder (Nakagawa et al. 2016). The research community has consistently identified neurodevelopmental disorders as being linked to specific temperament configurations (Johnson H. Johnson Mark H. et al. 2014). Several researchers also include activity level, attentional control, and impulsivity as temperament dimensions. According to the above conceptualizations, it should come as little surprise that the association between emotional problems and lack of self-regulatory mechanisms in children with ASD and parental stress is strong enough to warrant speculation that the disorder is perhaps better understood dimensionally.

In conclusion, the perplexing phenomenon of parenting stress in families of children with ASD requires a holistic approach to thoroughly evaluate the possible influence of multiple dimensions simultaneously. The age range and level of cognitive development of the individuals with ASD in the samples have been quite heterogeneous (Hastings et al. 2005b; Manning et al. 2011; Pozo and Sarriá 2014a, 2014b; Zaidman-Zait et al. 2014, 2018; Giovagnoli et al. 2015). However, so far, sporadic studies have investigated the predictors of parental stress in school children with ASD without intellectual disability (Bundy and Kunce 2009; Lee et al. 2009; Mori et al. 2009; Craig et al. 2016).

1.2. Hypothesis of the Present Study

Under the aegis of the above theoretical and research findings, the current study attempts to evaluate two different objectives. The first purpose of the study was to assess the complex relationship between parental stress and the severity of the ASD symptoms, behavioral difficulties, coping strategies, and social support. The second objective of the study was to evaluate the influence of behavioral difficulties within the sphere of temperament, coping strategies, and social support between ASD symptoms and parenting stress. We hypothesized that parenting stress would be positively correlated with ASD symptom severity, the temperament of the children, and coping strategies related to distraction and disengagement (Hypothesis 1). In addition, parenting stress would be negatively correlated with engagement and cognitive reframing coping strategies and social confidence and affective functional support (Hypothesis 2). Additionally, a harmonic balance among individuals and their environment is produced through a two-way interaction between inherent and temperamental traits and external experiences and circumstances, as well (Rothbart 2007).

2. Method

2.1. Participants

For the present study, 75 parents with at least one child with an autism diagnosis participated in the survey (M = 36.2, SD = 8.9). All the families lived in Greece and were recruited through special public schools and parental support groups. A member of the research team contacted the parents to explain the objectives of the study and request their collaboration. The children had received a clinical diagnosis of an autism spectrum condition in the psychiatry and child neurology services of public hospitals and medical centers in Greece, at ages ranging between 2 years and 11 months and 6 years (mean age of diagnosis = 4.69; SD = 1.67). The diagnosis was generally made using a multi-team approach, based on the Diagnostic and Statistical Manual of Mental Disorders (DSM-5; APA 2013) and the Autistic Diagnostic Interview—Revised (ADI-R; Rutter et al. 2006).

Finally, 18 class teachers with a mean age (years) M = 32.15 (SD = 3.62) working with the children in public primary special schools completed the Motor Behavior Checklist to rate the self-regulation skills of their students.

Inclusion criteria: (1) families with at least one child with an autism diagnosis; (2) families that live in Greece; (3) children with an autism diagnosis, either secondary or idiopathic autism; (4) children diagnosed from public hospitals and medical centers; (5) children in primary special public schools; (6) children older than 6 yearsand up 12 years (primary school); (7) parents that are participating in parental support groups.

Exclusion criteria: (1) children diagnosed with ASD from private practices and not from public hospitals and medical centers; (2) children not in primary special public schools; (3) children aged younger than 6 years or older than 12 years; (4) families not in parental support groups; (5) families not living in Greece.

2.2. Assessment Instruments

2.2.1. Parenting Stress Index–Short Form

Parenting Stress Index–Short Form (PSI-SF; Abidin 1995; Adapted to Greek) was used in this research. This scale was a self-report measure filled out by the parents. It contained 36 items distributed in three subscales of 12 items each, rated on a five-point Likert-type response scale. The first scale, parental distress, evaluated the distress experienced by parents due to personal factors, such as depression or conflict with a partner, or life restrictions due to the demands of childrearing in their role as parents (i.e., "Since having my child, I feel that I am almost never able to do things I like to do"). The second scale, parent–child dysfunctional interaction, provided information on the parents' feelings about the interactions with their child and the degree of frustration of the expectations and trust they have placed in their child (i.e., "Most times, I feel that my child does not like me and does not want to be close to me"). The third scale, difficult child, was designed to measure the parents' perceptions of their child's self-regulatory abilities (i.e., "My child seems to cry or fuss more often than most children"). The scale also provided a measure of total stress by adding up the scores on the 36 items, with a total score above 90 being clinically significant. The Cronbach's alpha internal consistency coefficients in our sample were parental distress (a = 0.91), dysfunctional parent–child interaction (a = 0.82), and difficult child (a = 0.90); this was similar to those obtained in other studies carried out in Spain (Diaz-Herrero et al. 2011). It is the most widely used instrument to evaluate stress in studies on ASD; in fact, it was utilized in 75% of the studies included in a recent systematic review (Barroso et al. 2018).

2.2.2. Motor Behavior Checklist for Children

In the present study, the self-regulation skills of ASD children were assessed using the Motor Behavior Checklist' (MBC) for children (Efstratopoulou et al. 2012b). The MBC checklist is a screening instrument designed to measure externalizing and internalizing behavioral symptoms of primary school-aged children. The instrument has been used in studies in Greece and has internal consistency (0.82 to 0.95), reproducibility (0.85 to 0.90), and interrater agreement (0.75 to 0.91) that have been checked in previous studies. More specifically, the MBC includes seven scales, which assess particular emotional and/or behavioral problems (i.e., rule breaking—7 items; hyperactivity/impulsivity—14 items; lack of attention—10 items; low energy—4 items; stereotyped behaviors—2 items; lack of social interaction—10 items; and lack of self-regulation—12 items). Many of these categories of behavioral problems can be observed in the form of both deficits and excesses in attention deficit hyperactivity disorder (ADHD) and autistic spectrum disorders (ASD) (Efstratopoulou et al. 2012a). The MBC should be completed by observing the child in a free-play situation or during physical education classes. The score is obtained through a 5-point Likert scale ranging from "never" (0) to "almost always" (4). Efstratopoulou et al. (2012a), evaluated the psychometric properties of the MBC: the coefficients of internal consistency (α) ranged from 0.82 to 0.95, reproducibility according to intraclass correlation

coefficients (ICC) ranged from 0.85 to 0.90, and concordance (also ICC) ranged from 0.75 to 0.91. These data suggest that the MBC for children is a homogeneous instrument in terms of content, with high stability and correlation (Efstratopoulou et al. 2012a). In addition, results from several studies on the psychometric properties of the checklist indicated that the MBC is a useful tool to discriminate between the core symptoms of ADHD, conduct disorder, and ASD (Efstratopoulou et al. 2012b). For the purposes of this study, mean scores on self-regulation items were calculated for all ASD children by their class teachers.

2.2.3. Child Autism Symptoms ASD Clinical Criteria from the DSM-5

The severity of the ASD symptoms was assessed through an interview between the parents and a clinical psychologist, focused on the seven diagnostic criteria for the disorder in accordance to DSM-5 (APA 2013). The first three were to evaluate socio-communicative impairments and the other four to rate repetitive behaviors and restricted interests. Through the interviews, the parents evaluated the severity of each criterion, using a 4-point Likert scale ranging from 0 to 3, where 0 represents "almost never", 1 "sometimes", 2 "often", and 3 "many times". Therefore, a higher score on the DSM-5 indicates greater severity of the ASD symptoms.

2.3. Procedure

All parents and teachers participated in the study voluntarily. Each parent received a file which contained: (a) the information letter concerning the objectives of the research, as well as the contact details of the research supervisor; (b) the form for completion of the demographic data; and (c) the Greek version of the Parenting Stress Index–Short Form questionnaire (PSI-SF; Abidin 1995). All participants were examined individually by the researcher who explained the procedure and conducted a small interview with them about the confidant and affective support they have in their everyday lives. Each parent had the opportunity to choose the place, as well as the time, to complete the questionnaires while at the same time, there was the possibility for clarifications, where this was necessary. No time limit was assigned for the completion of the questionnaire. Through the information letter, parents were encouraged to respond honestly, to ensure the reliability of the results. The participants gave written informed consent at the time of their visit, agreeing that their participation was voluntary and that they could withdraw at any time, without giving a reason and without cost. Due to the specific type of current research, demographic data such as age, gender, or occupation were selected. Since these are considered personal data, the European Union law that has existed since 28 May 2018 was applied.

According to the law, the use of sensitive personal data is allowed only due to research reasons. Therefore, the participants were informed accordingly, and they agreed that their personal data could be deleted from the web-database after a written request. This study obtained ethical approval and the participants' parents, after being informed about the objectives, signed the consent forms to participate in the study. They were fully aware that they could leave if they so desired. Next, an individual brief interview was administered by an accredited psychologist to confirm the diagnosis, record the social support for the family and complete the symptom severity list with the parent, in order to extract the necessary data on their children to carry out this study. Finally, all teachers completed the Greek Version of the Motor Behavior Checklist for Children (Efstratopoulou et al. 2012b).

2.4. Data Analyses

The statistical analyses for the current study were performed using the IBM SPSS Statistics software, Version 24.0 (Statistical Package for Social Science). To examine the relationship between self-regulation problems in ASD children and parental stress, Pearson correlations were calculated. The bivariate Pearson correlation produces a sample correlation coefficient, r, which measures the strength and direction of linear relationships between pairs of continuous variables. By extension, the Pearson correlation evaluates whether there is statistical evidence for a linear relationship among the same pairs of variables in

the population, represented by a population correlation coefficient, ϱ ("rho"). The Pearson correlation is a parametric measure.

In the second step of statistical analysis, in order to evaluate possible gender differences on a total parental stress scores, a one-way ANOVA was applied between mothers and fathers.

3. Results

With regard to the ASD children, most of them were boys (63%), and their ages ranged from 7 to 11, with a mean age of 8.59 (SD = 1.38). The children attended classes in special primary public schools, and they were receiving extra educational support of varying degrees (20 out of the 45 children were enrolled in communication and language classrooms).

In terms of parents and family characteristics, the parents' mean age (years) was 40.17 (SD = 4.82). With regard to their education level, 40 of the participants (53.3%) had obtained a university degree, 30 participants (40%) had obtained a high school diploma, and five participants (6.7%) had studies corresponding to primary education. The majority of the parents were employed (67%), whereas nine reported that they were unemployed at the moment, or they had never worked outside the house/family. With regard to their marital status, most of them were married—60 participants (80%)—and the rest (15) were separated/divorced (20%).

All demographic characteristics of parents are presented in Table 1.

Table 1. Parents' Demographic characteristics.

	N	%
Gender	75	
Male		27%
Female		83%
Marital status		
Married		80%
Divorced		14%
Single—No Family		6%
Educational Level		
University Level		53.3%
Secondary Level		40%
Primary/No Education		6.7%
Working Status		
Currently Working		67%
Unemployed or never worked in the past		33%

To test "Hypothesis 1" and "Hypothesis 2", Pearson correlation coefficients among research variables were estimated. The Pearson product–moment correlation analyses revealed the existence of significant associations between the ASD symptoms from the DSM-5, and the parenting stress total index ($p < 0.001$), Strengths and Difficulties Questionnaire (SDQ) ($p < 0.001$), engagement ($p < 0.001$), confidant support ($p < 0.001$), and affective support ($p < 0.05$). In addition, as shown in Table 2, the analyses revealed significant correlations between the parenting stress total index and the SDQ ($p < 0.001$), engagement ($p = 0.04$), and confidant support ($p = 0.05$). Self-regulation scores and scores on the separate subscales of the Parental Stress Index scale were calculated. Intercorrelations between variables are presented in Table 2.

Table 2. Intercorrelations between parental stress and self-regulation problems in ASD Children (N = 75).

Variable	Mean	SD	Parental Stress	Parent–Child Dysfunction	Difficult Child	Defense Responses	Total Parental Stress	Self-Regulation Scores in Children
Parental stress	25.2	4.8	–	0.486	0.394	**0.738** **	**0.704** **	0.297
Parent–Child Dysfunction	26.0	5.8	0.486	–	0.592 *	0.392	**0.651** **	0.292
Difficult child	25.0	5.8	0.394	0.592 *	–	0.623 *	**0.618** **	**0.761** **
Defense Responses	17.3	3.3	**0.738** **	0.392	0.623 *	–	**0.675** **	0.661 *
Total Parental Stress	87.0	17.2	**0.704** **	0.651*	**0.618** **	**0.675** **	–	**0.713** **
Self-regulation scores in children	29.7	5.7	0.297	0.292	0.661 *	0.349 **	0.713 **	–

Note: ASD = autism spectrum disorders, ** Correlation is significant at the 0.01 level (2-tailed), * Correlation is significant at the 0.05 level (2-tailed).

There were significant correlations between scores on self-regulation problems reported by the teachers and the total parental stress scores (r = 0.713) and the mean scores on the subscale difficult child (r = 0.761) reported by the parent.These results mean that the severity of self-regulation problems experienced by children with ASD influenced and increased parental stress.

In the second step of statistical analysis, research findings revealed that there were no statistical differences in total scores on stress between mothers and fathers for our sample F (2,72) = 10.951, $p < 0.001$. However, the Pearson correlation revealed that the educational level of parents was positively correlated with the level of parent–child dysfunction (r = 0.761) and the total stress exhibited by the parent (r = 0.654).

3.1. Interviews with Parents

A focus group of six parents (four mothers and two fathers) participated in structured interviews. A simple random sampling was used for the selection of the group, and every individual had an equal chance of being selected. A transcendental psychological phenomenology qualitative approach proposed by Moustakas (1994), in which each experience stands in its unique features and the phenomenon is introduced in a fresh complete description of thoughts, perceptions, and feelings, was used. The approach derives "*a textural description of the meanings and essences of the phenomenon, the constituents that comprise the experience in consciousness, from the vantage point of an open self*" (p. 35). This implies the subjectivity of the researcher and his/her role in presenting the importance of the phenomenon (Moustakas 1994). The interview recordings were transcribed and analyzed using Moustakas's (1994) phenomenological approach and Creswell and Poth's (2018) approach. First, the interview data were prepared for analysis by transcribing the audio and videotaped interviews. Significant statements or quotes from the transcripts that were essential for understanding the phenomenon were highlighted and coded (horizontalization) based on the main themes exploring the participants' experience (textural description) and how they experienced this phenomenon (structural description providing the themes with the retrieved quotes from the transcripts). Finally, the transcripts were sent to an external reviewer, who examined the themes and reflected on the validity and suitability of the themes to the purpose of the study and its questions.

3.1.1. Findings from the Interviews

The official ASD diagnosis evoked negative emotions of shock, blame, denial, and depression, leaving parents overwhelmed by the magnitude of the situation:

"I went into . . . severe depression . . . I used to get panic attacks (Liza)."

"I was thinking to hurt myself (John)."

Parents felt relieved as they understood their child's behavior and could move forward. Liza explained:

"I did feel relief. I understand. I know he is not a naughty child. The realization that parents were not to blame was liberating."

3.1.2. Parent's Anxiety and Body Exhaustion

Mothers reported that they felt nervous, and had bodily exhaustion. Anna commented that:

"He takes all my strength, efforts and make me easily nervous. I swear to God I do everything to accommodate his needs and ignore challenging behaviors but I am so tired to continue doing so."

Mothers are overwhelmed with the many responsibilities; Liza commented:

"I am so stressed having three children to look after them. I have no time to relax and think."

Parents' relationships with their neurotypical children were compromised, as they were mostly involved with the autistic child. Mothers expressed guilt about spending less time with their neurotypical children, but explained that neurotypical children understood their commitments:

"I cannot remember him at seven . . . eight . . . nine, I remember the sadness in his face, but I cannot remember anything [else]" (Liza). One neurotypical child reported to his parents that he felt "left out and unloved" as Nick explained.

In addition, the demands of a young child with ASD affected spousal relationships:

"Initially, we were blaming each other later, I was blaming myself. The family becomes dysfunctional; you blame your husband" (Hellen). A change to the couple relationship was unavoidable due to the investment of time with the child. There is no time together as a couple."

Difficulties relating to self-care activities were distressing, especially as parents were aware of their mortality:

"When he comes out of the bathroom . . . he does not care who is there. He will just run . . . now he is bigger (Helen). How long we will live and who will take care of her?" (Liza).

Sensory overstimulation resulted in meltdowns, which reinforced decisions to stay home:

"I cannot let him go to parties . . . if he gets a meltdown . . . people will not understand" (John).

Difficulties with communication were another stressor, as parents did not know what was wrong:

"It is not easy. Sometimes you just do not know . . . if he is hungry, if he needs to go to the toilet" (Nick).

Changes to routine generated stress for the checklist for autism spectrum disorder (CASD):

"She goes to school with the school transport; if the transport does not come, she would not want to go" (Nick).

Families had to ensure that routines were in place but sometimes had no control over the situation.

3.1.3. Using Distance Learning at Home: A Better Understanding of Children's Abilities

Parents valued the benefit of being close to their children during online sessions at home. They got to know their children's strengths and weaknesses. To Liza, online learning was:

> *"A blessing gift for her child to proceed to learn and be engaged with activities during the lockdown period at home."*

In addition, some of them were surprised by the abilities they had and never knew about before. Asma commented:

> *"He surprised me. My husband came to the room saw him interacting with the teacher on the screen and communicating on activities and he asked me surprisingly: do you know that he could do that?"*

However, not all experiences were good. Helen indicated that her child's behavior could not be controlled at home. He could not sit for more than 10 min in front of the computer or a desk to do his homework. She commented:

> *"He moves all the time and she cannot afford him to sit and listen all the time, . . . I invent my strategy to let him sit . . . . I sit and put him on my lap and then hug him with my legs and hands so that he stays in place and concentrates on what he is learning . . . but I get tired I did not pursue . . . It is difficult for the Autistic child with ADHD child to sit at all scheduled times; he gets bored easily even crying and gains nothing."*

Helen commented:

> *"I feel a lot of stress, body pain, and confused brain. I always feel guilty. If I worked with my son, I did not check on his other brothers or I did not check with the baby. I blame myself every day and get angry if I missed one of his assignments not finished. I feel I am running every single day from 5 a.m. to 8 p.m. I do not stop running . . . Running . . . . no break until I fall asleep of tiredness."*

It is worthy to note how the disability itself affects the whole family; it casts shadows of the usual stresses of accomplishing the learning process. The impulsivity of children adds another burden to the parents' worry and anxiety. Children with ADHD are usually prone to accidents, as mentioned by a mother's comment that her child fell from moving her brothers' Jeb car while they were on a picnic, running after his falling slippers. In addition to this, the challenge becomes worse when it is accompanied by another disability, such as ADHD or sensory problems. This was noticed from the replies of Helen, Liza, and Asma. Although with Asma the difficulties were not so obvious, this may have been because of her good economic status and level of education.

Some of the difficulties mentioned during the interviews were related to the child's disability, whereas others were related to school and the learning process, and others were related to the parents themselves (Lassoued et al. 2020). Children with ASD have unique characteristics that should be considered while learning online. These children should be taught according to their characteristics, having a continuous break between classes. Another suggestion by parents was that awareness workshops should be given to the societal community about Autism and the importance of taking into consideration the characteristics of included students. Financial aids and online learning resources such as computers should be devoted to parents who cannot afford them, and technical guidance should be provided to parents to improve their practical skills in supporting their children at home.

Parents' perceived anxiety was one of the findings of the study related to children's characteristics, as the severity of the behavioral problems affected the stress levels in parents. Thus, the whole family should be supported psychologically and socially to complete their role effectively and enjoy life. In addition, mothers with more than one child and greater responsibilities are more in need of this support.

4. Discussion

The study aimed to identify possible relationships between ASD children's severity of symptoms and several other factors known to play a significant role in the levels of stress among parents with children with developmental disabilities.

As Miranda et al. (2019) point out, although, during the last decade, an increase in the prevalence of parental stress in this population has been reported (Baio et al. 2018), nonetheless, scarce research exists which thoroughly explores the relationship between risk and protective factors and parental stress. According to Kiami and Goodgold (2017), parental stress of children with ASD has been found to reach clinically significant levels in 77% of the cases, and is greater than the stress of parents of children with typical development (Giovagnoli et al. 2015; Rao and Beidel 2009; Davis and Carter 2008). Moreover, it largely exceeds the parental stress of children with other neurodevelopmental disorders (e.g., ADHD, specific learning disorders, intellectual disabilities, etc.) (Gupta 2007; Hayes and Watson 2013; Watson et al. 2013; Craig et al. 2016; Barroso et al. 2018). In general, the findings of the current study were consistent with those reported among the relevant literature (Feldman and Werner 2002; Gray 2003; Hutton and Caron 2005; Mancil et al. 2009).

According to the overall results, the significant correlation values detected among most variables were largely expected. Data analysis indicated that the level of the child's functioning and his/her behavioral difficulties were significantly correlated with parental stress. There were significant correlations between scores on self-regulation problems reported from the teachers and the total parental stress scores and the mean scores on the subscale difficult child reported by the parent, a research finding that confirmed both of our hypotheses.

Additionally, most of the findings coincided with the majority of previous relevant literature. More specifically, the significant association found between higher levels of parental stress and the increase in the core symptomatology of ASD was in line with other studies, which emphasized a strong correlation between the severity of autism symptoms and higher levels of parental stress (Miranda et al. 2019; Ben-Sasson et al. 2013; Tomeny 2017; Bitsika and Sharpley 2017, etc.). The same finding validated the fact that positive and problem-focused strategies strongly correlate with less severe symptoms of ASD, which is also in line with the findings of Kiami and Goodgold (2017), Lai et al. (2015), Obeid and Daou (2014), and Benson (2010).

The literature revealed that each developmental disability, due to its unique behavioral characteristics, was itself a source of continuous anxiety for parents, as was the parent's perception of the actual disability (Baio et al. 2018; Barroso et al. 2018; Stanojevic et al. 2017; Benson 2010) etc. Especially regarding ASD, problematic behaviors may include physical aggression, self-injury, property destruction, stereotyped behaviors, tantrums, etc. As a result, children with ASD are often highly disruptive to the classroom, home environments and the community (Horner et al. 2002; Efstratopoulou et al. 2012a). All of these behaviors have been directly related to parental stress (Kiami and Goodgold 2017; Barroso et al. 2018; Mancil et al. 2009). Zaidman-Zait et al. (2017), stated that mothers experienced lower levels of stress when they utilized more active coping strategies and relied less on disengaged coping strategies, either at the time of diagnosis or overtime.

On the other hand, parental stress was negatively correlated with the level of social functional support reported by the parents. This finding validates previous studies arguing that social support can significantly reduce the anxiety and distress experienced by parents raising a child with ASD (Lindsey and Barry 2018). In particular, Boyd (2002) found that a common coping strategy that decreased parental stress in this subgroup was contact with family members and parents of other children with autism. However, in the cases in which autism symptoms were more intense, it was found that parents were more reluctant to share their intimate feelings with other people. As a result, in these cases, the problem-focused coping strategies tended to reduce their effectiveness. These parents also declared that they received less empathy and caring from social sources.

Previous literature investigating the coping strategies used by parents of children with ASD to deal with various daily stressors has revealed conflicting findings, especially with regard to the long-term effectiveness of specific strategies often addressed by parents (e.g., social withdrawal, separating the child with ASD from his/her siblings, etc.). These findings can serve as useful guidelines to researchers and offer practical advice for intervention planning for practitioners working with families dealing with ASD, who often exhibit increased levels of stress (Boyd 2002).

Furthermore, correlation analyses also revealed that parental stress was positively associated with children's high scores on the MBC subscale (rated by their teachers), which confirmed the high prevalence and intensity of ASD symptoms. This finding is in agreement with other studies that have reported strong correlations among ASD children's problematic behaviors (rated with the use of several behavioral screening scales by parents and/or teachers), and the severity of ASD symptoms and parental stress (Posserud et al. 2018; Helland and Helland 2017). The strong link between parental stress and the emotional/behavioral disorder (EBD) of children with ASD found in this study has also been highlighted in other studies (Yorke et al. 2018; Barroso et al. 2018). This positive association of stress seems to exceed the value for the relationship between ASD symptom severity and EBD, which was also evident in the correlation analyses of this study (Miranda et al. 2019).

The findings from this study revealed a strong correlation between parental stress and children's high scores on the MBC subscale. The higher the score on the MBC, the higher the level of parental stress. In addition, this correlation was also beneficial in terms of the issue of cross-informant agreement (i.e., the agreement between different informants' multiple sources of information, e.g., parents, teachers, children or youth themselves) about a child's overall functioning in different settings for the MBC (Achenbach et al. 2017). Little research is available in the area of cross-informant agreement (especially among parents and teachers). More specifically, even though, in this study, children were rated with the MBC only by their teachers, nevertheless, in this special online learning situation (due to the lockdown), parents and teachers had the unique opportunity to simultaneously observe children's behaviors in exactly the same context.

As a result of the cross-informant agreement mentioned above, a unique common view of teachers and parents of children with ASD could be used as a valuable source for a better understanding of how to intervene to alter problematic behaviors of ASD children in both school and home settings. This will provide parents with better, more effective coping skills to deal with these behaviors. Likewise, the positive relationship between ASD symptoms and behavioral problems was confirmed, as consistently reported in the literature (Helland and Helland 2017; Posserud et al. 2018).

The ASD symptoms, as expected, were significantly and negatively associated with engagement coping and with social support suggesting that mothers who perceive the autism symptoms of their children with greater intensity tend to reduce their problem-focused coping strategies think they can communicate their intimate feelings to other people less, and receive fewer demonstrations of caring and empathy. Regarding parenting stress, the correlations generally support our hypotheses about the expected relationships. In addition to correlating with ASD symptoms, parenting stress presented a positive association with behavioral problems, exceeding the value for the relationship between ASD symptoms and behavioral problems.

Parents must receive help through family-centered supportive services that offer counseling, to decrease their stress levels by using appropriate coping strategies and other resources. Brief interventions that include stress management, details about specific behavioral impairments, and principles of behavior management within a set of components (information on autism, strategies for teaching new skills, improving social interaction and communication, service availability, family and community responses to autism) have shown their effectiveness in reducing parenting stress and improving family life (Kasari et al. 2015). Furthermore, according to an emerging body of research, mindfulness-based

interventions may help reduce parenting stress in mothers who have children with ASD (Conner and White 2014).

Limitations and Future Research

Even though the present study provided some innovative conceptualizations concerning the parental stress of children with ASD and the role of ASD symptom severity, self-regulation and coping strategies on parental stress, the study has some limitations. Addressing those limitations may lead to further research in this field. First, since the sample was relatively small, future studies could include larger samples to generalize findings.

In addition, future research should include cross informant ratings with the MBC (and other similar screening tools) of both parents and teachers of children (and adolescents) with ASD and other developmental disorders during distance learning situations. This would further validate the cross-informant agreement of such instruments and lead to better intervention strategies and techniques for ASD children's learning, which will derive from the common experiences of parents and teachers.

Apart from parental educational levels, other factors could also be included in future research, such as socioeconomic status, everyday life conditions, etc. Finally, findings from similar studies can have many practical implications, in terms of awareness planning, special education training workshops and communication skills training for parents of children with ASD and other developmental difficulties. This could reduce the anxiety levels reported by parents of ASD and other developmental disorders and make them feel more confident to support their children.

5. Conclusions

As the literature suggests, the issue of parental stress and their psychological adaptation in the extremely difficult situation of rearing a child with ASD is a very complex variable, which depends upon a combination of both risk and protective factors. These include the personal characteristics and behavioral profile of the child, the severity of the core symptomatology of ASD, the frequency and severity of emotional and behavioral difficulties manifested, the family's positive and negative coping strategies, as well as the level of social and/or other types of support (e.g., educational) parents receive, especially in stressful situations.

In line with previous research, the present study also revealed that both ASD symptomatology and EBD were highly correlated with high levels of anxiety in parents, whereas engagement coping, sufficient or high educational level and social functional support were factors, which negatively correlated with parental stress. Likewise, findings confirmed that the prevalence of less severe ASD symptoms and better self-regulation skills for the children werepositively correlated with coping strategies used by parents and, consequently, with a reduced anxiety level. In addition, results confirmed the mediating role of EBD, parents' coping strategies and social functional support in the association between parental stress and the symptom severity of children with ASD, also reported in previous studies.

Hence, eliminating the stressors parents face in raising children with ASD does not seem possible. Instead, improving parental coping and resilience should be the objective in helping family functioning when there are new and ongoing challenges.

Finally, the present study points out the need to promote parents' coping orientation and the application of behavioral strategies with their children to help them handle the immense impact of stress. Suggestions to support families with children with ASD aim at the development of strategies for the improvement of the self-regulation skills of nonverbal children with ASD. This may be particularly important in terms of reducing parental stress for families of children with autism and other developmental disabilities.

Author Contributions: Conceptualization, M.E., M.S., S.G., and E.B.; methodology, M.E., M.S., S.G., and E.B.; software, M.E., M.S., and E.B.; validation, M.E., M.S., and E.B.; formal analysis, M.E., and M.S.; investigation, M.E., M.S., S.G., and E.B.; resources, M.E., M.S., S.G., and E.B.; data curation, M.E.,

and M.S.; writing—original draft preparation, M.E., M.S., S.G., and E.B.; writing—review and editing, M.E., M.S., S.G., and E.B.; visualization, M.E., M.S., S.G., and E.B.; supervision, M.E., M.S., and E.B.; project administration, M.E., M.S., and E.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: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of United Arab Emirates University (UAEU) (protocol code: ERS_2021_7335 and date of approval: 20 June 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the patient(s) to publish this paper.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy issues. The data are not publicly available due to privacy.

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

Abbreviations

The following abbreviations are used in this manuscript:

ABA	Applied Behavior Analysis
ADHD	Attention Deficit Hyperactivity Disorder
ADI-R	Autistic Diagnostic Interview-Revised
ASD	Autism Spectrum Disorder
CASD	Checklist for Autism Spectrum Disorder
DSM-5	Diagnostic and Statistical Manual of Mental Disorder-5
EBD	Emotional/Behavioral Disorder
ICC	Intraclass Correlation Coefficients
IQ	Intellectual Quotient
LEAP	Learning Experiences: An Alternative Program for Preschoolers and Parents
MBC	Motor Behaviour Checklist for children
PECS	Picture Exchange Communication System
PSI-SF	Parenting Stress Index–Short Form
TEACCH	Treatment and Education of Autistic and related Communication handicapped Children
SDQ	Strengths and Difficulties Questionnaire
SPSS	Statistical Package for Social Science

References

Abidin, R. Richard. 1995. *Parenting Stress Index: Professional Manual*, 3rd ed. Odessa: Psychological Assessment Resources.

Achenbach, Thomas M., Masha Y. Ivanova, and Leslie A. Rescorla. 2017. Empirically based assessment and taxonomy of psychopathology for ages 1½–90+ years: Developmental, multi-informant, and multicultural findings. *Comprehensive Psychiatry* 79: 4–18. [CrossRef]

American Psychiatric Association (APA). 2013. *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. (DSM 5). Washington, DC: American Psychiatric Association.

Baio, Jon, Lisa Wiggins, Deborah L. Christensen, Matthew J. Maenner, Julie Daniels, Zachary Warren, Margaret Kurzius-Spencer, Walter Zahorodny, Cordelia Robinson Rosenberg, Tiffany White, and et al. 2018. Prevalence of autism spectrum disorder among children aged 8 years. Autism and developmental disabilities monitoring network, 11 Sites, United States 2014. *Surveillance Summaries* 67: 1–23. [CrossRef] [PubMed]

Barroso, Nicole E., Lucybel Mendez, Paulo A. Graziano, and Daniel M. Bagner. 2018. Parenting stress through the lens of different clinical groups: A systematic review & meta-analysis. *Journal of Abnormal Child Psychology* 46: 449–61. [CrossRef] [PubMed]

Ben-Itzchak, Esther, Linda R. Watson, and Ditza A. Zachor. 2014. Cognitive ability is associated with different outcome trajectories in autism spectrum disorders. *Journal of Autism and Developmental Disorders* 44: 2221–29. [CrossRef]

Ben-Sasson, Ayelet, Timothy W. Soto, Frances de L. Martínez-Pedraza, and Alice S. Carter. 2013. Early sensory over- responsivity in toddlers with autism spectrum disorder as a predictor of family impairment and parenting stress. *Journal of Child Psychology and Psychiatry* 54: 846–53. [CrossRef] [PubMed]

Benson, Paul R. 2010. Coping, distress, and well-being in mothers of children with autism. *Research in Autism Spectrum Disorders* 4: 217–28. [CrossRef]

Bitsika, Vicki, and Christopher F. Sharpley. 2017. The association between autism spectrum disorder symptoms in high-functioning male adolescent and their mother's anxiety and depression. *Journal of Developmental and Physical Disabilities* 9: 461–73. [CrossRef]

Boyd, Brian A. 2002. Examining the relationship between stress and lack of social support in mothers of children with autism. *Focus on Autism and Other Developmental Disabilities* 17: 208–15. [CrossRef]

Bundy, Myra Beth, and Linda J. Kunce. 2009. Parenting stress and high functioning children with autism. *International Journal on Disability and Human Development* 8: 401–10. [CrossRef]

Centers for Disease Control and Prevention (CDC). 2021. Treatment and Intervention Services for Autism Spectrum Disorder. Available online: https://www.cdc.gov/ncbddd/autism/treatment.html (accessed on 23 September 2021).

Conner, Caitlin M., and Susan W. White. 2014. Stress n mothers of children with autism: Trait mindfulness as a protective factor. *Research in Autism Spectrum Disorders* 8: 617–24. [CrossRef]

Craig, Francesco, Francesca F. Operto, Andrea De Giacomo, Lucia Margari, Alessandro Frolli, Massimiliano Conson, Sara Ivagnes, Marianna Monaco, and Francesco Margari. 2016. Parenting stress among parents of children with neurodevelopmental disorders. *Psychiatry Research* 242: 121–29. [CrossRef]

Creswell, Josh W., and Cheryl N. Poth. 2018. *Qualitative Inquiry and Research Design: Choosing among Five Approaches*, 4th ed. Thousand Oaks: Sage Publications.

Davis, Naomi O., and Alice S. Carter. 2008. Parenting stress in mothers and fathers of toddlers with autism spectrum disorders: Associations with child characteristics. *Journal of Autism and Developmental Disorders* 38: 1278–91. [CrossRef] [PubMed]

Diaz-Herrero, Ángela, José Antonio López-Pina, Julio Pérez-López, Alfredo G. Brito de la Nuez, and María Teresa Martínez-Fuentes. 2011. Validity of the parenting stress index-short form in a sample of Spanish fathers. *Spanish Journal of Psychology* 14: 990–97. [CrossRef] [PubMed]

Efstratopoulou, Maria, Rianne Janssen, and Johan Simons. 2012a. Agreement among physical educators, teachers and parents on children's behaviors: A multitrait-multimethod design approach. *Research in Developmental Disabilities* 33: 1343–51. [CrossRef]

Efstratopoulou, Maria, Rianne Janssen, and Johan Simons. 2012b. Differentiating children with attention- deficit/hyperactivity disorder, conduct disorder, learning disabilities and autistic spectrum disorders by means of their motor behavior characteristics. *Research in Developmental Disabilities* 33: 196–204. [CrossRef] [PubMed]

Feldman, Maurice A., and Shannon E. Werner. 2002. Collateral effects of behavioral parents training on families of children with developmental disabilities and behavior disorders. *Behavioral Interventions* 17: 75–83. [CrossRef]

Foster, Anne, Debbie Rude, and Caroline Grannan. 2012. Preparing parents to advocate for a child with autism. *The Phi Delta Kappan* 94: 16–20. [CrossRef]

Giovagnoli, Giulia, Valentina Postorino Laura M. Fatta, Veronica Sanges, Lavinia De Peppo, Lia Vassena, Paola De Rose, Stefano Vicari, and Luigi Mazzonea. 2015. Behavioural and emotional profile and parental stress in preschool children with autism spectrum disorder. *Research in Developmental Disabilities* 45–46: 411–21. [CrossRef]

Gray, David E. 2003. Gender and coping: The parents of children with high functioning autism. *Social Science & Medicine (1982)* 56: 631–42. [CrossRef]

Gupta, Vidya Bhushan. 2007. Comparison of parenting stress in different developmental disabilities. *Journal of Developmental and Physical Disabilities* 19: 417–25. [CrossRef]

Hastings, Richard P. 2003. Child behavior problems and partner mental health as correlates of stress in mothers and fathers of children with autism. *Journal of Intellectual Disability Research* 47: 231–37. [CrossRef]

Hastings, Richard P., Hanna Kovshoff, Nicholas J. Ward, Francesca degli Espinosa, Tony Brown, and Bob Remington. 2005a. Systems analysis of stress and positive perceptions in mothers and fathers of pre-school children with autism. *Journal of Autism and Developmental Disorders* 35: 635–44. [CrossRef] [PubMed]

Hastings, Richard P., Hanna Kovshoff, Tony Brown, Nicholas J. Ward, Francesca degli Espinoza, and Bob Remington. 2005b. Coping strategies in mothers and fathers of preschool and school-age children with autism. *Autism* 9: 377–91. [CrossRef] [PubMed]

Hayes, Stephanie A., and Shelley L. Watson. 2013. The impact of parenting stress: A meta-analysis of studies comparing the experience of parenting stress in parents of children with and without autism spectrum disorder. *Journal of Autism and Developmental Disorders* 43: 629–42. [CrossRef] [PubMed]

Helland, Wenche Andersen, and Turid Helland. 2017. Emotional and behavioural needs in children with specific language impairment and in children with autism spectrum disorder: The importance of pragmatic language impairment. *Research in Developmental Disabilities* 70: 33–39. [CrossRef] [PubMed]

Horner, Robert H., Edward G. Carr, Phillip S. Strain, Anne W. Todd, and Holly K. Reed. 2002. Problem Behavior Interventions for Young Children with Autism: A Research Synthesis. *Journal of Autism and Developmental Disorders* 32: 423–46. [CrossRef] [PubMed]

Hutton, Adam M., and Sandra L. Caron. 2005. Experiences of Families With Children With Autism in Rural New England. *Focus on Autism and Other Developmental Disabilities* 20: 180–89. [CrossRef]

Johnson H. Johnson Mark H., Teodora Gliga, Emily Jones, and Tony Charman. 2014. Annual research review: Infant development, autism, and ADHD: Early pathways to emerging disorders. *Journal of Child Psychology and Psychiatry* 56: 228–47. [CrossRef] [PubMed]

Kasari, Connie, Amanda Gulsrud, Tanya Paparella, Gerhard Hellemann, and Kathleen Berry. 2015. Randomized comparative efficacy study of parent-mediated interventions for toddlers with autism. *Journal of Consulting and Clinical Psychology* 83: 554–63. [CrossRef]

Kiami, Sheri R., and Shelley Goodgold. 2017. Support needs and coping strategies as predictors of stress level among mothers of children with autism spectrum disorder. *Autism Research and Treatment* 2017: 1–10. [CrossRef] [PubMed]

Lai, Wei Wei, Tze Jui Goh, Tian P. S. Oei, and Min Sung. 2015. Coping and well-being in parents of children with autism spectrum disorders (ASD). *Journal of Autism and Developmental Disorders* 45: 2582–93. [CrossRef] [PubMed]

Lassoued, Zohra, Mohammed Alhendawi, and Raed Bashitialshaaer. 2020. An exploratory study of the obstacles for achieving quality in distance learning during the COVID-19 pandemic. *Education Sciences* 10: 232. [CrossRef]

Lee, Ching-Fang, Fang-Ming Hwang, Chwen-Jen Chen, and Li-Yin Chien. 2009. The interrelationships among parenting stress and quality of life of the caregiver and preschool child with very low birth weight. *Family Community Health* 32: 228–37. [CrossRef]

Lindsey, Rebecca A., and Tammy D. Barry. 2018. Protective factors against distress for caregivers of a child with autism spectrum disorder. *Journal of Autism and Developmental Disorders* 48: 1092–107. [CrossRef] [PubMed]

Mancil, Richmond G., Brian A. Boyd, and Pena Bedesem. 2009. Parental Stress and Autism: Are There Useful Coping Strategies? *Education and Training in Developmental Disabilities* 44: 523–37.

Manning, Margaret M., Laurel Wainwright, and Jillian Bennett. 2011. The double ABCX model of adaptation in racially diverse families with a school-age child with autism. *Journal of Autism and Developmental Disorders* 41: 320–31. [CrossRef]

Martins, Rosa, Inês Bonito, Ana Andrade, Carlos Albuquerque, and Claudia Chaves. 2015. The impact of the diagnosis of autism in parents of children. *Procedia-Social and Behavioral Sciences* 171: 121–25. [CrossRef]

Matson, Johnny L., and Alison M. Kozlowski. 2011. The increasing prevalence of autism spectrum disorders. *Research in Autism Spectrum Disorders* 5: 418–25. [CrossRef]

Meadan, Hedda, James W. Halle, and Aaron T. Ebata. 2010. Families with children who have autism spectrum disorders: Stress and support. *Exceptional Children* 77: 7–36. [CrossRef]

Miranda, Ana, Mira Alvaro, Carmen Berenguer, Belen Rosello, and Inmaculada Baixauli. 2019. Parenting Stress in Mothers of Children with Autism without Intellectual Disability. Mediation of Behavioral Problems and Coping Strategies. *Frontiers in Psychology* 10: 464. [CrossRef] [PubMed]

Mori, Kyoko, Takeshi Ujiie, Anna Smith, and Patricia Howlin. 2009. Parental stress associated with caring for children with Asperger's syndrome or autism. *Pediatrics International* 51: 364–70. [CrossRef] [PubMed]

Moustakas, Clark. 1994. *Phenomenological Research Methods*. Thousand Oaks: Sage.

Nakagawa, Atsuko, Masune Sukigara, Taishi Miyachi, and Akio Nakai. 2016. Relations between temperament, sensory processing, and motor coordination in 3-year-old children. *Frontier in Psychology* 7: 623–34. [CrossRef] [PubMed]

Obeid, Rita, and Nidal Daou. 2014. The effects of coping style, social support, and behavioral problems on the well-being of mothers of children with autism spectrum disorders in Lebanon. *Research in Autism Spectrum Disorders* 10: 59–70. [CrossRef]

Poppi, Kristi, Julia Jones, and Nicola Botting. 2019. Childhood autism in the UK and Greece: A cross-national study of progress in different intervention contexts. *International Journal of Developmental Disabilities* 65: 162–74. [CrossRef]

Posserud, M., Mari Hysing, Wenche Andersen Helland, Christopher Gillberg, and Astri J. Lundervold. 2018. Autism traits: The importance of co-morbid problems for impairment and contact with services. Data from the Bergen child study. *Research in Developmental Disabilities* 72: 275–83. [CrossRef] [PubMed]

Pozo, Pilar, and Sarria Sarriá. 2014a. A global model of stress in parents of children with autism spectrum disorders (ASD). *Anales de Psicologia* 30: 180–91. [CrossRef]

Pozo, Pilar, and Sarria Sarriá. 2014b. Prediction of stress in mothers of children with autism spectrum disorders. *Spanish Journal of Psychology* 17: E6. [CrossRef]

Rao, Patricia A., and Deborah C. Beidel. 2009. The impact of children with high-functioning autism on parental stress, sibling adjustment and family functioning. *Behavior Modification* 33: 437–51. [CrossRef] [PubMed]

Rothbart, Mary K. 2007. Temperament, development, and personality. *Current Directions in Psychological Science* 16: 207–12. [CrossRef]

Rutter, Michael, Ann Le Couteur, and Catherine Lord. 2006. *ADI-R. EntrevistaClínica Para el Diagnóstico del Autismo-Revisada*. Madrid: TEA Ediciones.

Sharabi, Adi, and Dafna Marom-Golan. 2018. Social support, education levels, and parents' involvement: A comparison between mothers and fathers of young children with autism spectrum disorder. *Topics in Early Childhood Special Education* 38: 54–64. [CrossRef]

Stanojevic, Ninaa, Vanjab Nenadović, Saškaa Fatić, and Miodragb Stokic. 2017. Exploring factors ´ of stress level in parents of children with autistic spectrum disorder. *Specijalna Edukacijai Rehabilitacija* 16: 445–63. [CrossRef]

Tincani, Matt, Maia Bloomfield Cucchiarra, S. Kenneth Thurman, Mark R. Snyder, and Catherine M. McCarthy. 2014. Evaluating NRC's recommendations for educating children with autism a decade later. *Child and Youth Care Forum* 43: 315–37. [CrossRef]

Tomeny, Theodore S. 2017. Parenting stress as an indirect pathway to mental health concerns among mothers of children with autism spectrum disorder. *Autism* 21: 907–11. [CrossRef] [PubMed]

Watson, Shelley L., Kelly D. Coons, and Stephanie A. Hayes. 2013. Autism spectrum disorder and fetal alcohol spectrum disorder. Part I: A comparison of parenting stress. *Journal of Intellectual and Developmental Disability* 38: 95–104. [CrossRef] [PubMed]

Yorke, Isabel, Pippa White, Amelia Weston, Monica Rafla, Tony Charman, and Emily Simonof. 2018. The association between emotional and behavioral problems in children with autism spectrum disorder and psychological distress in their parents: A systematic review and meta-analysis. *Journal of Autism and Developmental Disorders* 48: 3393–415. [CrossRef]

Zaidman-Zait, Anat, Pat Mirenda, Eric Duku, Peter Szatmari, Stelios Georgiades, Joanne Volden, Lonnie Zwaigenbaum, Tracy Vaillancourt, Susan Bryson, Isabel Smith, and et al. 2014. Examination of bidirectional relationships between parent stress and two types of problem behavior in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders* 44: 1908–17. [CrossRef] [PubMed]

Zaidman-Zait, Anat, Pat Mirenda, Eric Duku, Tracy Vaillancourt, Isabel M Smith, Peter Szatmari, Susan Bryson, Eric Fombonne, Joanne Volden, Charlotte Waddell, and et al. 2017. Impact of personal and social resources on parenting stress in mothers of children with autism spectrum disorder. *Autism* 21: 155–66. [CrossRef] [PubMed]

Zaidman-Zait, Anat, Pat Mirenda, Peter Szatmari, Eric Duku, Isabel M. Smith, Tracy Vaillancourt, Joanne Volden, Charlotte Waddell, Teresa Bennett, Lonnie Zwaigenbaum, and et al. 2018. Profiles of social and coping resources in families of children with autism spectrum disorder: Relations to parent and child outcomes. *Journal of Autism and Developmental Disorders* 48: 2064–76. [CrossRef] [PubMed]

Journal of
Intelligence

MDPI

Article

Reconsidering the Use of the Mindset Assessment Profile in Educational Contexts

Alexander P. Burgoyne [1,*] and Brooke N. Macnamara [2]

[1] School of Psychology, Georgia Institute of Technology, 654 Cherry St., Atlanta, GA 30332, USA
[2] Department of Psychological Sciences, Case Western Reserve University, 11220 Bellflower Road, Cleveland, OH 44106-7123, USA; bnm24@case.edu
* Correspondence: burgoyn4@gmail.com

Abstract: The Mindset Assessment Profile is a popular questionnaire purportedly designed to measure mindset—an individual's belief in whether intelligence is malleable or stable. Despite its widespread use, the questionnaire appears to assess an individual's need for cognition and goal orientation more than mindset. We assessed the reliability, construct validity, and factor structure of the Mindset Assessment Profile in a sample of 992 undergraduates. The reliability of the Mindset Assessment Profile was questionable (α = .63) and significantly lower than the reliability of the Implicit Theories of Intelligence Questionnaire (α = .94), an established measure of mindset. The Mindset Assessment Profile also lacked convergent and discriminant validity. Overall scores on the Mindset Assessment Profile correlated significantly more strongly with need for cognition than with mindset. Item-level analyses supported this finding: most items correlated weakly or not at all with mindset, and correlated significantly more strongly with need for cognition and learning goal orientation. Exploratory factor analysis indicated that three factors were underlying scores on the Mindset Assessment Profile: need for cognition, mindset, and performance goal orientation. Based on its questionable reliability and poor construct validity, we do not recommend that researchers and educators use the Mindset Assessment Profile to measure mindset.

Keywords: mindset; implicit theories; mindset assessment profile; validity; reliability

Citation: Burgoyne, Alexander P., and Brooke N. Macnamara. 2021. Reconsidering the Use of the Mindset Assessment Profile in Educational Contexts. *Journal of Intelligence* 9: 39. https://doi.org/10.3390/jintelligence9030039

Received: 28 May 2021
Accepted: 30 July 2021
Published: 4 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Reconsidering the Use of the Mindset Assessment Profile in Educational Contexts

Mindset refers to people's beliefs about the nature of their abilities. People with a growth mindset believe that attributes such as intelligence can be developed, whereas people with a fixed mindset believe that these attributes are stable. Some researchers have argued that a growth mindset is beneficial for academic achievement, on the premise that students with a growth mindset will pursue challenges and be resilient to setbacks (Blackwell et al. 2007). However, recent empirical evaluations have revealed only weak evidence for most of mindset's premises, including its relationships with goal orientation, pursuit of challenges, resilience to setbacks, and academic achievement (Burgoyne et al. 2020; Li and Bates 2019; Payne et al. 2007; Sisk et al. 2018). Nevertheless, educators frequently use growth mindset interventions to encourage students to adopt a growth mindset in an effort to improve their academic performance (Boaler 2013; Sisk et al. 2018).

Mindset interventions are a booming industry in the educational sector. For example, the for-profit company Mindset Works has sold growth mindset interventions products to parents, teachers, and schools for over a decade (https://www.mindsetworks.com/programs/ (accessed on 6 April 2021)). Included in these intervention programs is an eight-item measure—the Mindset Assessment Profile—which is used as a diagnostic tool to determine whether students have a growth or fixed mindset; that is, whether they believe intelligence is malleable or stable (Hall 2016; Thomas 2018). The Mindset Assessment Profile questionnaire is also provided on the company's website, where visitors are encouraged to "Take the Mindset Assessment to Learn More About Your Mindset."

(http://blog.mindsetworks.com/what-s-my-mindset (accessed on 6 April 2021)). As a result, researchers and educators frequently administer the Mindset Assessment Profile to measure mindset in educational contexts (see, e.g., Bedford 2017; Cartwright and Hallar 2018; Hall 2016; Lim et al. 2020; Neufville 2019; Saia 2017; Thomas 2018; Wakefield 2019; Wolferd 2020). Often, the Mindset Assessment Profile is administered before and after an intervention to test whether it altered students' mindsets. A change from pre- to post-intervention is taken as evidence that the intervention was successful (Bedford 2017; Cartwright and Hallar 2018; Saia 2017; Thomas 2018; Wolferd 2020), a point we return to in the Discussion.

Given the Mindset Assessment Profile's extensive use among researchers, educators, and students, the psychometric qualities of this scale have practical significance. Although Mindset Works describes the Mindset Assessment Profile as a "diagnostic tool drawn from research-validated measures" (http://blog.mindsetworks.com/what-s-my-mindset (accessed on 6 April 2021)), they do not provide any information on the reliability, construct validity, or factor structure of the scale.

2. Present Study

We noticed that many items in the Mindset Assessment Profile appeared to be tapping constructs other than mindset, namely goal orientation—one's drive to master new material and demonstrate competency (Elliot and Church 1997), and need for cognition—one's tendency to engage in and enjoy thinking (Cacioppo and Petty 1982). If this is the case, students, teachers, and parents may have a misconstrued understanding of their mindset based on their Mindset Assessment Profile scores.

According to mindset theory, goal orientations are related to one's mindset of intelligence, but they are distinct constructs (Dweck and Leggett 1988). That is, individuals with a growth mindset are hypothesized to endorse learning goals, reflecting a desire to acquire new skills, whereas people with a fixed mindset are hypothesized to endorse performance goals, reflecting a desire to prove their abilities (or not demonstrate a lack of ability). Despite these claims, however, evidence suggests that mindset is only weakly related to goal orientation. For example, in a sample of 438 undergraduate students, Burgoyne et al. (2020) found that mindset was weakly correlated with learning goal orientation ($r = .10$) and performance goal orientation ($r = -.11$), and Payne et al. (2007) found correlations of a similar magnitude in meta-analytic work. In light of their results, Payne et al. (2007) concluded that the relationship between mindset and goal orientation had been overstated by proponents of mindset theory: "Contrary to Dweck's (1986) perspective, the effect sizes were very small, providing little evidence for Dweck's (1986) view that implicit theories are the primary underlying antecedent of GO [goal orientation]" (p. 140).

Need for cognition, on the other hand, is a relatively unexplored construct within mindset's nomological network. At a conceptual level, one might expect that individuals with more of a growth mindset would rate higher on need for cognition. That is, mindset theory would likely predict that individuals with a growth mindset would enjoy mental challenges, such as thinking about complex problems, on the basis that they might learn from them. The empirical evidence for this relationship is scarce, but suggests a weak correlation. For example, Birney et al. (2018) found that growth mindset correlated $r = .12$ with need for cognition in a sample of 142 experienced business managers.

A related concern is that researchers using the Mindset Assessment Profile may make inaccurate assumptions about the relationship between mindset and measured outcomes (e.g., academic achievement) if the Mindset Assessment Profile includes items measuring other non-mindset constructs. For instance, if Mindset Assessment Profile scores are contaminated by the inclusion of goal orientation items, the observed relationship between Mindset Assessment Profile scores and goal orientation will be exaggerated. As another example, if goal orientation and need for cognition are stronger predictors of academic achievement than mindset is, then the relationship between Mindset Assessment Profile

scores and academic achievement will be artificially inflated due to the inclusion of items tapping these constructs.

Analyses

The purpose of this study was to assess the internal consistency reliability, construct validity, and factor structure of the Mindset Assessment Profile as a measure of mindset. We estimated the internal reliability of the Mindset Assessment Profile by computing Cronbach's alpha (α; Cronbach 1951; see George and Mallery 2003, for rules of thumb for interpreting Cronbach's alpha) and McDonald's omega coefficient (ω; McDonald 1999; McNeish 2018; Zinbarg et al. 2005). Cronbach's alpha tests for consistency among items within a measure, and McDonald's omega indicates the proportion of variance in the scale scores accounted for by a single factor (Zinbarg et al. 2005).

Construct validity is the degree to which a measure's variance is attributable to variance in the construct it is intended to measure rather than some other factor (O'Leary-Kelly and Vokurka 1998). Construct validity is evaluated in terms of convergent and discriminant validity. Convergent validity is the degree to which different measures designed to assess the same construct correlate with one another (Cunningham et al. 2001): measures of the same construct should be strongly related. We tested for convergent validity by correlating scores on the Mindset Assessment Profile with a well-established measure of mindset, Dweck's (2000) Implicit Theories of Intelligence Questionnaire, which has been shown to have sound psychometric properties (Dweck 2000). Discriminant validity, on the other hand, refers to the extent to which measures designed to assess different constructs correlate with one another (Campbell and Fiske 1959). Compared with two measures assessing the same construct, measures designed to assess different constructs should be more weakly correlated. We tested for discriminant validity by correlating scores on the Mindset Assessment Profile with measures of need for cognition and goal orientation.

Finally, we conducted an exploratory factor analysis on the items in the Mindset Assessment Profile to assess its factor structure. If all or most items load well onto a single factor, this suggests the Mindset Assessment Profile is measuring a single personality construct. If items load better on multiple factors, this suggests multiple personality constructs are underlying scores on this measure.

3. Method

Methods were pre-registered at https://osf.io/N82F4/.

3.1. Participants

The participants were 998 undergraduate students at Michigan State University, ranging in age from 18 to 31 (M = 19.73, SD = 1.48). Approximately 63% of the participants were female. Around 38% of the participants were in their first year of college, 28% were in their second year, 21% were in their third year, and the remaining 13% were in their fourth or fifth year. Six participants were excluded because they did not reach the end of the survey, leaving a final sample of 992 participants. Missing data (<1% of cases) were handled using listwise deletion on an analysis-by-analysis basis. With 992 participants, we had 89% power to detect significant correlations of $r \geq .10$ (Faul et al. 2007). All participants provided informed consent and received partial course credit for their participation in the study.

3.2. Measures

Demographics. Participants were asked to report their age, year in college, and gender.

Mindset Assessment Profile. Participants responded to the eight items in the Mindset Assessment Profile taken from the Mindset Works website (http://blog.mindsetworks.com/what-s-my-mindset (accessed on 6 April 2021)) using a six-point Likert scale ranging from "Disagree a lot" to "Agree a lot." The items are listed in order of administration: (1) "No matter how much intelligence you have, you can always change it a good deal";

(2) "You can learn new things, but you cannot really change your basic level of intelligence"; (3) "I like my work best when it makes me think hard"; (4) "I like my work best when I can do it really well without too much trouble"; (5) "I like work that I'll learn from even if I make a lot of mistakes"; (6) "I like my work best when I can do it perfectly without any mistakes"; (7) "When something is hard, it just makes me want to work more on it, not less"; and (8) "To tell the truth, when I work hard, it makes me feel as though I'm not very smart." Even numbered items were reverse scored. The final score was the mean response to the items.

Mindset. Participants completed Dweck's (2000) Implicit Theories of Intelligence Questionnaire as a measure of mindset. Participants responded to eight items using a seven-point Likert scale, rating the degree to which they agreed or disagreed with each statement: (1) "You have a certain amount of intelligence, and you can't really do much to change it"; (2) "No matter who you are, you can significantly change your intelligence level"; (3) "Your intelligence is something about you that you can't change very much"; (4) "You can always substantially change how intelligent you are"; (5) "To be honest, you can't really change how intelligent you are"; (6) "No matter how much intelligence you have you can always change it quite a bit"; (7) "You can learn new things, but you can't really change your basic intelligence"; and (8) "You can change even your basic intelligence level considerably." Odd numbered items were reverse scored. The response options ranged from "Strongly agree" to "Strongly disagree." The final score was the mean response to the items. Higher scores on this measure correspond to more of a growth mindset, reflecting the belief that intelligence is malleable. Lower scores correspond to more of a fixed mindset, reflecting the belief that intelligence is stable.

Need for Cognition. Participants completed Cacioppo et al.'s (1984) Need for Cognition Questionnaire. Participants responded to eighteen items using a seven-point Likert scale, rating the degree to which they agreed or disagreed with statements such as "I would prefer complex to simple problems" and "Thinking is not my idea of fun" (reverse scored). The response options ranged from "Strongly agree" to "Strongly disagree." The final score was the mean response to the items. Higher scores correspond to greater need for cognition.

Goal Orientation. Participants completed an adapted version of Elliot and Church's (1997) Goal Orientation Questionnaire. Participants responded to sixteen items using a seven-point Likert scale ranging from "Disagree a lot" to "Agree a lot." This questionnaire assesses three goal orientations: learning goal orientation, performance approach goal orientation, and performance avoidance goal orientation. Participants rated the degree to which they agreed or disagreed with learning goal statements such as "I want to learn as much as possible," performance approach goal statements such as "I strive to demonstrate my ability relative to others," and performance avoidance goal statements such as "I worry about the possibility of performing poorly." The final score for each goal orientation was the mean response to the items. Higher scores correspond to greater endorsement of each goal orientation.

Procedure. Participants completed the questionnaires online on Qualtrics. Participants were first presented with the three-item demographic questionnaire. The order of the remaining questionnaires was randomized across participants to control for potential order effects.

4. Results

Analyses were preregistered at https://osf.io/N82F4/ and were conducted using SPSS. Data are openly available at https://osf.io/N82F4/.

4.1. Reliability of the Mindset Assessment Profile

Descriptive statistics are presented in Table 1 and distributions are presented in Figure 1. Of the questionnaires administered to participants, the Mindset Assessment Profile had the lowest reliability (α = .63, ω = .61). Table A1 in the Appendix A presents the inter-item correlations for the Mindset Assessment Profile. The correlations between

the items varied in absolute magnitude from $r = .05$ to $r = .49$. Compared to the Mindset Assessment Profile, the Implicit Theories of Intelligence Questionnaire ("Mindset" in Table 1) had excellent reliability ($\alpha = .94$, $\omega = .94$). Indeed, the Mindset Assessment Profile had a significantly lower Cronbach's alpha reliability estimate than the Implicit Theories of Intelligence Questionnaire, $t = 38.89$, $p < .001$ (Abd-El-Fattah and Hassan 2011). The other questionnaires had acceptable to good reliability, ranging from $\alpha = .72$ to $\alpha = .87$ (ω also ranged from .72 to .87).

Table 1. Descriptive statistics.

Measure	Items	*N*	*M*	*SD*	Skew	Kurtosis	α	ω
Mindset Assessment Profile	8	992	3.70	0.62	0.11	−0.02	.63	.61
Mindset	8	992	4.72	1.26	−0.31	−0.28	.94	.94
Need for cognition	18	991	4.37	0.79	0.08	0.46	.87	.87
Learning goal orientation	5	991	5.48	0.84	−0.45	0.30	.78	.78
Performance approach goal orientation	6	991	4.96	1.07	−0.42	−0.07	.85	.86
Performance avoidance goal orientation	5	991	5.39	0.98	−0.62	0.15	.72	.72

Note. "Mindset" refers to the Implicit Theories of Intelligence Questionnaire.

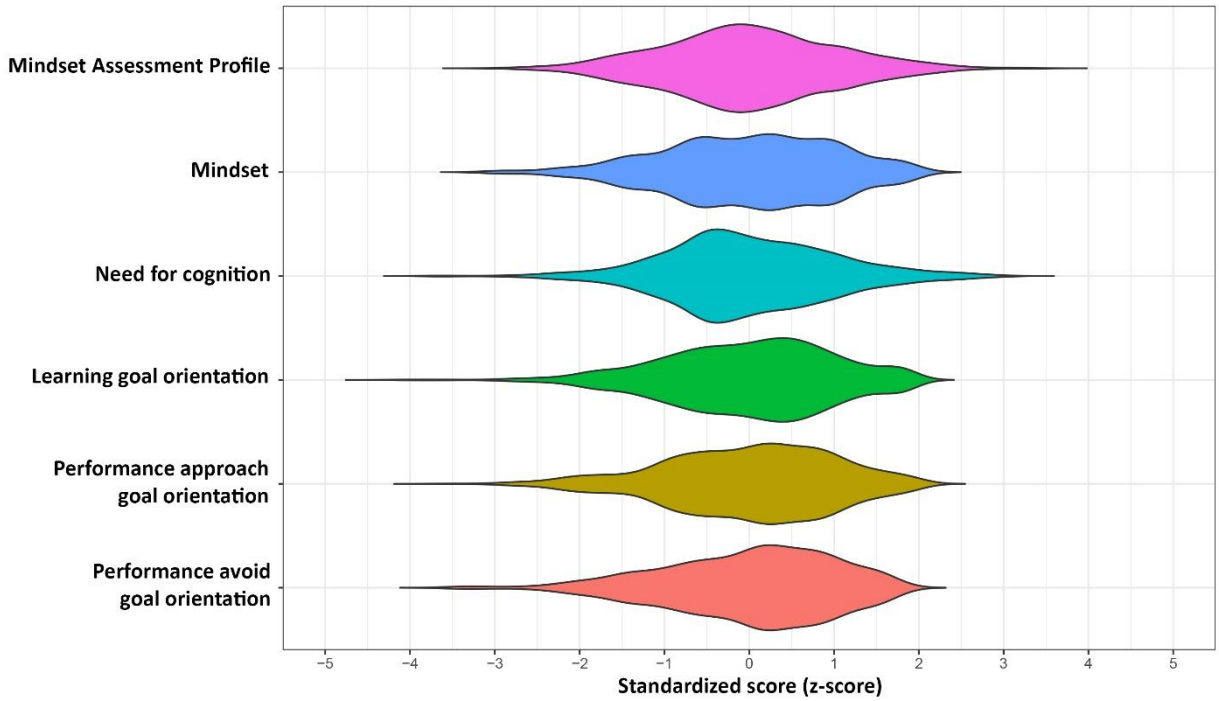

Figure 1. Distribution of standardized scores (i.e., *z*-scores) on each measure. *Note:* "Mindset" refers to the Implicit Theories of Intelligence Questionnaire.

4.2. Construct Validity of the Mindset Assessment Profile

Correlations between measures are presented in Table 2. Scatterplots depicting the relationships between the Mindset Assessment Profile and the other measures (upper row) and between the Implicit Theories of Intelligence Questionnaire and the other measures (bottom row) are presented in Figure 2. Scores on the Mindset Assessment Profile correlated most strongly with need for cognition ($r = .59$, 95% CI [.55, .63], $p < .001$), followed by mindset ($r = .50$, 95% CI [.45, .55], $p < .001$) and learning goal orientation ($r = .48$, 95% CI [.43, .53], $p < .001$). Steiger's (1980) test for the difference between dependent correlations revealed that the correlation between the Mindset Assessment Profile and need for cognition was significantly stronger than the correlation between the Mindset Assessment Profile and mindset, $z = 2.87$, $p = .004$. This indicates that the Mindset Assessment Profile lacks construct validity due to poor discriminant validity. Scores on the Mindset Assessment

Profile were more closely related to need for cognition than mindset. For comparison, mindset as measured by the Implicit Theories of Intelligence Questionnaire correlated only weakly with need for cognition (r = .18, 95% CI [.12, .24], p < .001); see Figure 2.

As an additional test of convergent and discriminant validity, we computed correlations between each of the items in the Mindset Assessment Profile and the other personality measures.[1] The purpose of this analysis was to understand which items in the Mindset Assessment Profile correlated more strongly with non-mindset constructs than with mindset.

As shown in Table 3, most of the items in the Mindset Assessment Profile correlated more strongly with need for cognition and learning goal orientation than with mindset. Only items one and two correlated strongly with mindset (r = .70, 95% CI [.67, .73], p < .001 and r = .71, 95% CI [.68, .74], p < .001, respectively). This is not surprising, as the wording of these items is nearly identical to the wording of items six and seven in the Implicit Theories of Intelligence Questionnaire.

By contrast, items three, four, five, six, seven, and eight from the Mindset Assessment Profile correlated significantly more strongly with need for cognition (average r = .38) than with mindset (average r = .12); Steiger's test for the difference between dependent correlations revealed zs ranging from 3.32 to 11.96, all ps < .001. Items three, four, five, six, seven, and eight from the Mindset Assessment Profile also correlated significantly more strongly with learning goal orientation (average r = .29) than with mindset (average r = .12); Steiger's zs ranged from 2.07 to 9.90, all ps < .02. Items four and six from the Mindset Assessment Profile did not correlate significantly with mindset (rs = −.01, 95% CIs [−.07, .05], ps > .65).

These results indicate that most of the items in the Mindset Assessment Profile lack both convergent and discriminant validity. Most items correlated weakly or not at all with mindset, and correlated significantly more strongly with need for cognition and learning goal orientation than with mindset. A correlation matrix with the items from all the personality scales is presented on the Open Science Framework: https://osf.io/N82F4/.

Table 2. Correlation matrix.

Personality Measure	1	2	3	4	5
(1) Mindset Assessment Profile	—				
(2) Mindset	**.50**	—			
(3) Need for cognition	**.59**	**.18**	—		
(4) Learning goal orientation	**.48**	**.21**	**.58**	—	
(5) Performance approach goal orientation	−.03	**−.09**	.06	**.27**	—
(6) Performance avoid goal orientation	**−.15**	.03	**−.12**	**.20**	**.38**

Note. Listwise n = 990. Correlation coefficients in bold are statistically significant at p < .05.

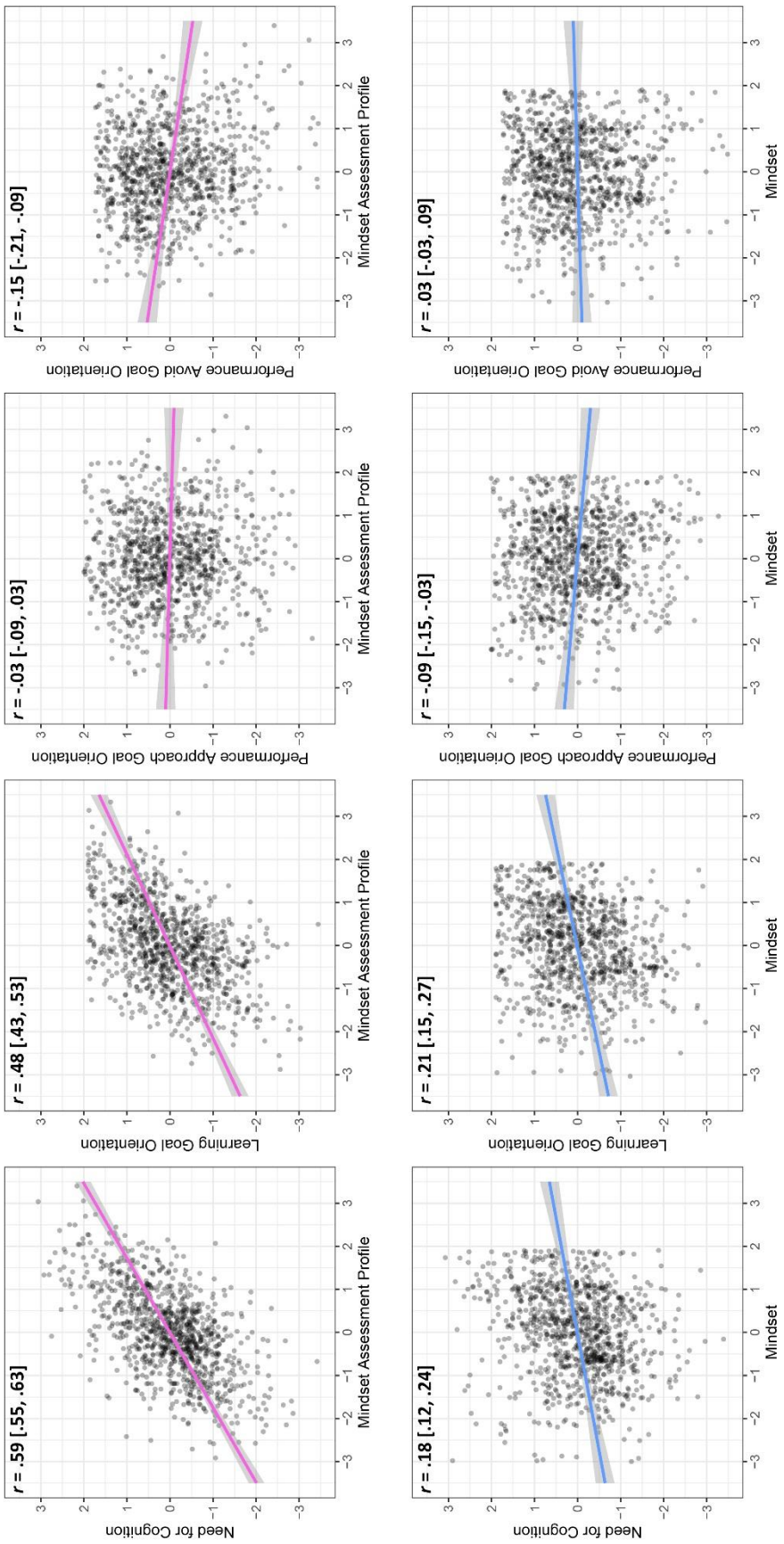

Figure 2. Scatterplots depicting correlations between Mindset Assessment Profile scores and other measures (top row) and between mindset and other measures (bottom row). *Note.* "Mindset" refers to the Implicit Theories of Intelligence Questionnaire. All scores are standardized (i.e., *z*-scores).

Table 3. Correlations between items in the Mindset Assessment Profile and personality measures.

Mindset Assessment Profile Item	Mindset	Need for Cognition	Performance Approach	Performance Avoidance	Learning Goal
(1) No matter how much intelligence you have, you can always change it a good deal.	**.70**	.10	-.02	.04	.17
(2) You can learn new things, but you cannot really change your basic level of intelligence.	**.71**	.14	−.06	.00	.13
(3) I like my work best when it makes me think hard.	.14	**.57**	.13	.01	.50
(4) I like my work best when I can do it really well without too much trouble.	−.01	**.32**	−.04	−.13	.15
(5) I like work that I'll learn from even if I make a lot of mistakes.	.22	**.38**	−.02	−.06	.34
(6) I like my work best when I can do it perfectly without any mistakes.	−.01	**.22**	−.14	−.21	.09
(7) When something is hard, it just makes me want to work more on it, not less.	.12	**.43**	.06	−.13	.39
(8) To tell the truth, when I work hard, it makes me feel as though I'm not very smart.	.20	**.33**	.02	−.12	.28

Note. Listwise n = 971. The correlation coefficient in bold is the strongest correlation for each item. $|rs| \geq .07$ are statistically significant at $p < .05$.

4.3. Factor Structure of the Mindset Assessment Profile

Finally, we conducted an exploratory factor analysis on the items in the Mindset Assessment Profile to determine whether a single personality factor or multiple personality factors were underlying scores on this measure. We used principal axis factoring with promax rotation to allow extracted factors to correlate, and extracted factors with Eigenvalues ≥1.0.

As shown in Table 4, three factors emerged from the exploratory factor analysis of the Mindset Assessment Profile. Items three, five, and seven had high loadings on the first factor; items one and two had high loadings on the second factor; and items four and six had high loadings on the third factor. Item eight did not load highly on any factor. Although subject to interpretation, the first factor appears to represent need for cognition, the second factor appears to represent mindset, and the third factor appears to represent performance goal orientation. These results suggest that the Mindset Assessment Profile is not a unidimensional measure of mindset, but rather that three factors underlie scores on this measure.

Table 4. Exploratory factor analysis of the Mindset Assessment Profile.

Mindset Assessment Profile Item	Factor 1	Factor 2	Factor 3
(1) No matter how much intelligence you have, you can always change it a good deal.	.21	**.56**	−.22
(2) You can learn new things, but you cannot really change your basic level of intelligence.	−.12	**.87**	.12
(3) I like my work best when it makes me think hard.	**.69**	−.04	.05
(4) I like my work best when I can do it really well without too much trouble.	.03	.00	**.72**
(5) I like work that I'll learn from even if I make a lot of mistakes.	**.62**	.02	−.06
(6) I like my work best when I can do it perfectly without any mistakes.	.06	.00	**.62**
7) When something is hard, it just makes me want to work more on it, not less.	**.56**	−.01	.09
(8) To tell the truth, when I work hard, it makes me feel as though I'm not very smart.	**.25**	.16	.09
Eigenvalue	2.28	1.53	1.15

Note. Listwise n = 973. The coefficient in bold is the strongest factor loading for each item. Correlations between factors: Factor 1 with Factor 2 (r = .26); Factor 1 with Factor 3 (r = .25), Factor 2 with Factor 3 (r = .00).

5. Discussion

We assessed the reliability, construct validity, and factor structure of the Mindset Assessment Profile in a sample of 992 undergraduate students. The internal reliability of the Mindset Assessment Profile (α = .63) was significantly lower than that of the Implicit

Theories of Intelligence Questionnaire (α = .94), which had excellent reliability. Both of these measures consist of eight items and were ostensibly designed to measure mindset.

Further, the Mindset Assessment Profile lacked construct validity as a measure of mindset. Overall scores on the Mindset Assessment Profile correlated significantly more strongly with need for cognition than with mindset. Item-level analyses supported this finding, revealing that six of eight items in the Mindset Assessment Profile correlated more strongly with both need for cognition ($\bar{r}$ = .38) and learning goal orientation ($\bar{r}$ = .29) than with mindset ($\bar{r}$ = .12). Only two of eight items from the Mindset Assessment Profile correlated strongly with mindset (*r*s = .70 and .71) as measured by the Implicit Theories of Intelligence Questionnaire. These items are nearly identical to items from the Implicit Theories of Intelligence Questionnaire. Finally, two of the eight items in the Mindset Assessment Profile had no association with mindset as measured by the Implicit Theories of Intelligence Questionnaire (*r*s = −.01, *p*s > .65).

Exploratory factor analysis revealed that three factors were underlying scores on the Mindset Assessment Profile. These factors appeared to represent need for cognition, mindset, and performance goal orientation. This corroborates the previous results by showing that the Mindset Assessment Profile is not a unidimensional measure of mindset.

The Mindset Assessment Profile is marketed as a measure of mindset. That is, students are encouraged to use the Mindset Assessment Profile to "assess their mindsets" on the Mindset Works website (http://blog.mindsetworks.com/what-s-my-mindset (accessed on 6 April 2021)). After completing the questionnaire, they are emailed a description of their "current mindset." Regardless of their results, they are directed to a webpage that sells growth mindset interventions ranging in cost from $20 per student to $7500 per school (https://www.mindsetworks.com/programs/ (accessed on 6 April 2021)).

Perhaps of greater concern, the Mindset Assessment Profile is included as a diagnostic tool in some of Mindset Works' growth mindset intervention programs. If the Mindset Assessment Profile is administered before and after a mindset intervention, change scores might be taken as evidence that an intervention successfully altered a student's mindset, when this effect would be more accurately described as a change in need for cognition.

This tendency to misconstrue Mindset Assessment Profile scores is not uncommon. As a case in point, Bedford (2017) administered the Mindset Assessment Profile to secondary school students before and after a growth mindset intervention "to evaluate the success of the growth mindset interventions" (p. 433). When Bedford (2017) found significantly different Mindset Assessment Profile scores following the intervention, this was interpreted as evidence that the "interventions put in place were successful in changing mindset towards a growth mindset" (p. 436). As another example, Wolferd (2020) recently administered the Mindset Assessment Profile to elementary school children to evaluate the effects of Mindset Works' Brainology program. Based on their scores on the Mindset Assessment Profile following the intervention, they reported that "female students in the experimental group displayed a significant, positive change in mindset" (p. 49) suggesting that "Brainology was an effective intervention for female students" (Wolferd 2020, p. 51). The research presented herein suggests that the Mindset Assessment Profile is more a measure of need for cognition than mindset, and that its description as a "mindset" assessment has led to misunderstandings about the efficacy of mindset interventions.

Relatedly, Lim et al. (2020) recently used the Mindset Assessment Profile as a diagnostic tool to assess college students in a work-study program. Based on their scores on the Mindset Assessment Profile, students were categorized into those with a "growth mindset" and those with a "fixed or unsure mindset" (p. 111). They found that "growth mindset" students' work-study supervisors rated them more highly on problem solving and decision-making than students not categorized as having a "growth mindset." However, based on our research, a more accurate conclusion would be that students with higher need for cognition are more likely to be rated higher on problem solving and decision-making than students lower in need for cognition.

In sum, despite the Mindset Assessment Profile's stated purpose as a mindset assessment and diagnostic tool, our results indicate that it is a poor measure of mindset. We recommend researchers avoid using the Mindset Assessment Profile as a measure of mindset or as a diagnostic tool in educational contexts.

Author Contributions: Conceptualization, A.P.B. and B.N.M.; methodology, A.P.B. and B.N.M.; software, A.P.B.; formal analysis, A.P.B.; investigation, A.P.B.; resources, A.P.B.; data curation, A.P.B.; writing—original draft preparation, A.P.B. and B.N.M.; writing—review and editing, A.P.B. and B.N.M.; visualization, A.P.B.; project administration, A.P.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 research was approved by the Institutional Review Board at Michigan State University on 6 June 2016 (IRB numbers: x16-753e; i051494).

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

Data Availability Statement: Data and preregistration are provided at the following Open Science Framework link: https://osf.io/n82f4/?view_only=f236b81a4d434711aff5db4b26f319a8.

Conflicts of Interest: The authors declare that they have no conflict of interest to report for the present manuscript.

Appendix A

Table A1. Correlation matrix with all items from the Mindset Assessment Profile.

Mindset Assessment Profile Item	1	2	3	4	5	6	7
(1) No matter how much intelligence you have, you can always change it a good deal.	—						
(2) You can learn new things, but you cannot really change your basic level of intelligence.	**.49**	—					
(3) I like my work best when it makes me think hard.	**.16**	**.07**	—				
(4) I like my work best when I can do it really well without too much trouble.	**−.09**	**.07**	**.20**	—			
(5) I like work that I'll learn from even if I make a lot of mistakes.	**.23**	**.08**	**.41**	.05	—		
(6) I like my work best when I can do it perfectly without any mistakes.	**−.10**	**.07**	**.15**	**.47**	**.15**	**—**	
(7) When something is hard, it just makes me want to work more on it, not less.	**.16**	**.08**	**.41**	**.20**	**.34**	**.16**	—
(8) To tell the truth, when I work hard, it makes me feel as though I'm not very smart.	**.12**	**.20**	**.25**	**.11**	**.20**	**.12**	**.16**

Note. Listwise $n = 973$. Correlations in bold are statistically significant at $p < .05$.

Notes

[1] This analysis was not pre-registered.

References

Abd-El-Fattah, Sabry M., and Hala K. Hassan. 2011. Dependent-alpha calculator: Testing the differences between dependent coefficients Alpha. *Journal of Applied Quantitative Methods* 11: 59–61.

Bedford, Susannah. 2017. Growth mindset and motivation: A study into secondary school science learning. *Research Papers in Education* 32: 424–43. [CrossRef]

Birney, Damian P., Jens F. Beckmann, Nadin Beckmann, Kit S. Double, and Karen Whittingham. 2018. Moderators of learning and performance trajectories in microworld simulations: Too soon to give up on intellect!? *Intelligence* 68: 128–40. [CrossRef]

Blackwell, Lisa S., Kali H. Trzesniewski, and Carol Sorich Dweck. 2007. Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development* 78: 246–63. [CrossRef]

Boaler, Jo. 2013. Ability and mathematics: The mindset revolution that is reshaping education. *Forum* 55: 143–52. [CrossRef]

Burgoyne, Alexander P., David Z. Hambrick, and Brooke N. Macnamara. 2020. How firm are the foundations of mind-set theory? The claims appear stronger than the evidence. *Psychological Science* 31: 258–67. [CrossRef] [PubMed]

Cacioppo, John T., and Richard E. Petty. 1982. The need for cognition. *Journal of Personality and Social Psychology* 42: 116–31. [CrossRef]

Cacioppo, John T., Richard E. Petty, and Chuan Feng Kao. 1984. The efficient assessment of need for cognition. *Journal of Personality Assessment* 48: 306–7. [CrossRef] [PubMed]

Campbell, Donald T., and Donald W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56: 81–105. [CrossRef] [PubMed]

Cartwright, T. J., and B. Hallar. 2018. Taking risks with a growth mindset: Long-term influence of an elementary pre-service after school science practicum. *International Journal of Science Education* 40: 348–70. [CrossRef]

Cronbach, Lee J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334. [CrossRef]

Cunningham, William A., Kristopher J. Preacher, and Mahzarin R. Banaji. 2001. Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science* 12: 163–70. [CrossRef] [PubMed]

Dweck, Carol S. 1986. Motivational processes affecting learning. *American Psychologist* 10: 1040–48. [CrossRef]

Dweck, Carol S. 2000. *Self-Theories: Their Role in Motivation, Personality, and Development*. East Sussex: Psychology Press.

Dweck, Carol S., and Ellen L. Leggett. 1988. A social-cognitive approach to motivation and personality. *Psychological Review* 95: 256–73. [CrossRef]

Elliot, Andrew J., and Marcy A. Church. 1997. A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology* 72: 218–32. [CrossRef]

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39: 175–91. [CrossRef] [PubMed]

George, Darren, and Paul Mallery. 2003. *SPSS for Windows Step by Step: A Simple Guide and Reference*, 11.0 update 4th ed. Boston: Allyn & Bacon.

Hall, Charity. 2016. Mindset and Instructional Practices in the Elementary Mathematics Classroom. (4121.). Master's Thesis, School of Education Student Capstone Theses and Dissertation, Saint Paul, MN, USA. Available online: https://digitalcommons.hamline.edu/hse_all/4121 (accessed on 6 April 2021).

Li, Yue, and Timothy C. Bates. 2019. You can't change your basic ability, but you work at things, and that's how we get hard things done: Testing the role of growth mindset on response to setbacks, educational attainment, and cognitive ability. *Journal of Experimental Psychology: General* 148: 1640–55. [CrossRef] [PubMed]

Lim, Sok Mui, Yong Lim Foo, May-Fung Yeo, Chelsea Yu Xian Chan, and Han Tong Loh. 2020. Integrated work study program: Students' growth mindset and perception of change in work-related skills. *International Journal of Work-Integrated Learning* 21: 103–15.

McDonald, Roderick P. 1999. *Test Theory: A Unified Treatment*. Mahwah: Lawrence Erlbaum.

McNeish, Daniel. 2018. Thanks coefficient alpha, we'll take it from here. *Psychological Methods* 23: 412–33. [CrossRef] [PubMed]

Neufville, Merica E. 2019. Perspectives of Mathematically Proficient Black High School Students with a History of Underachievement in Mathematics. Ph.D. thesis, St. John Fisher College, Rochester, NY, USA. Available online: https://fisherpub.sjfc.edu/education_etd/420/ (accessed on 6 April 2021).

O'Leary-Kelly, Scott W., and Robert J. Vokurka. 1998. The empirical assessment of construct validity. *Journal of Operations Management* 16: 387–405. [CrossRef]

Payne, Stephanie C., Satoris S. Youngcourt, and J. Matthew Beaubien. 2007. A meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology* 92: 128–50. [CrossRef] [PubMed]

Saia, Katherine. 2017. Impact of Mindset on Literacy: What Happens to Literacy Skills When a Growth Mindset Is Taught to First Graders. (2346). Master's Thesis, Rowan University, Glassboro, NJ, USA. Available online: https://rdw.rowan.edu/etd/2346 (accessed on 6 April 2021).

Sisk, Victoria F., Alexander P. Burgoyne, Jingze Sun, Jennifer L. Butler, and Brooke N. Macnamara. 2018. To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science* 29: 549–71. [CrossRef] [PubMed]

Steiger, James H. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin* 87: 245–51. [CrossRef]

Thomas, Phillip. 2018. The Impact of Teaching Growth Mindset on Archery Skill Achievement: An Action Research Study. Ph.D. thesis, University of South Carolina, Columbia, SC, USA. Available online: https://scholarcommons.sc.edu/etd/5101 (accessed on 6 April 2021).

Wakefield, Maria C. 2019. An Examination of Mindset and Academic Growth of Middle School Science Students. Ph.D. thesis, Regent University, Virginia Beach, VA, USA. Available online: https://search.proquest.com/docview/2248012392?pq-origsite=gscholar&fromopenview=true (accessed on 6 April 2021).

Wolferd, Jaclyn N. 2020. A Growth Mindset Intervention with Elementary-Age Children. Ph.D. thesis, Alfred University, Alfred, NY, USA. Available online: https://aura.alfred.edu/bitstream/handle/10829/23571/Wolferd_Jaclyn_2020.pdf (accessed on 6 April 2021).

Zinbarg, Richard E., William Revelle, Iftah Yovel, and Wen Li. 2005. Cronbach's α, Revelle's β, and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika* 70: 123–33. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
www.mdpi.com

Journal of Intelligence Editorial Office
E-mail: jintelligence@mdpi.com
www.mdpi.com/journal/jintelligence

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-0551-8