



education sciences

Scientific Reasoning in Science Education

From Global Measures to Fine- Grained Descriptions of Students' Competencies

Edited by

Moritz Krell, Andreas Vorholzer and Andreas Nehring

Printed Edition of the Special Issue Published in *Education Sciences*

**Scientific Reasoning in Science
Education: From Global Measures to
Fine-Grained Descriptions of
Students' Competencies**

Scientific Reasoning in Science Education: From Global Measures to Fine-Grained Descriptions of Students' Competencies

Editors

Moritz Krell

Andreas Vorholzer

Andreas Nehring

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin



Editors

Moritz Krell

Biology Education

IPN - Leibniz Institute for

Science and Mathematics

Education

Kiel

Germany

Andreas Vorholzer

Department Educational

Sciences

Technical University

of Munich

Munich

Germany

Andreas Nehring

Institute for Science

Education

Leibniz Universität Hannover

Hannover

Germany

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 *Education Sciences* (ISSN 2227-7102) (available at: www.mdpi.com/journal/education/special_issues/Scientific_Reasoning_Science_Education).

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

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. <i>Journal Name</i> Year , <i>Volume Number</i> , Page Range.
--

ISBN 978-3-0365-4548-6 (Hbk)

ISBN 978-3-0365-4547-9 (PDF)

© 2022 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license, which allows users to download, copy and build upon published articles, as long as the author and publisher are properly credited, which ensures maximum dissemination and a wider impact of our publications.

The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons license CC BY-NC-ND.

Contents

Preface to "Scientific Reasoning in Science Education: From Global Measures to Fine-Grained Descriptions of Students' Competencies" vii

Moritz Krell, Andreas Vorholzer and Andreas Nehring

Scientific Reasoning in Science Education: From Global Measures to Fine-Grained Descriptions of Students' Competencies

Reprinted from: *Educ. Sci.* **2022**, *12*, 97, doi:10.3390/educsci12020097 1

Besim Enes Bicak, Cornelia Eleonore Borchert and Kerstin Höner

Measuring and Fostering Preservice Chemistry Teachers' Scientific Reasoning Competency

Reprinted from: *Educ. Sci.* **2021**, *11*, 496, doi:10.3390/educsci11090496 9

Dagmar Hilfert-Rüppell, Monique Meier, Daniel Horn and Kerstin Höner

Professional Knowledge and Self-Efficacy Expectations of Pre-Service Teachers Regarding Scientific Reasoning and Diagnostics

Reprinted from: *Educ. Sci.* **2021**, *11*, 629, doi:10.3390/educsci11100629 33

Samia Khan and Moritz Krell

Patterns of Scientific Reasoning Skills among Pre-Service Science Teachers: A Latent Class Analysis

Reprinted from: *Educ. Sci.* **2021**, *11*, 647, doi:10.3390/educsci11100647 65

Moritz Krell, Samia Khan and Jan van Driel

Analyzing Cognitive Demands of a Scientific Reasoning Test Using the Linear Logistic Test Model (LLTM)

Reprinted from: *Educ. Sci.* **2021**, *11*, 472, doi:10.3390/educsci11090472 75

Daniela Mahler, Denise Bock and Till Bruckermann

Preservice Biology Teachers' Scientific Reasoning Skills and Beliefs about Nature of Science: How Do They Develop and Is There a Mutual Relationship during the Development?

Reprinted from: *Educ. Sci.* **2021**, *11*, 558, doi:10.3390/educsci11090558 91

Erika Schlatter, Ard W. Lazonder, Inge Molenaar and Noortje Janssen

Individual Differences in Children's Scientific Reasoning

Reprinted from: *Educ. Sci.* **2021**, *11*, 471, doi:10.3390/educsci11090471 111

Anna Beniermann, Laurens Mecklenburg and Annette Upmeier zu Belzen

Reasoning on Controversial Science Issues in Science Education and Science Communication

Reprinted from: *Educ. Sci.* **2021**, *11*, 522, doi:10.3390/educsci11090522 125

Valeria M. Cabello, Patricia M. Moreira and Paulina Griñó Morales

Elementary Students' Reasoning in Drawn Explanations Based on a Scientific Theory

Reprinted from: *Educ. Sci.* **2021**, *11*, 581, doi:10.3390/educsci11100581 143

Vanessa Lang, Christine Eckert, Franziska Perels, Christopher W. M. Kay and Johann Seibert

A Novel Modelling Process in Chemistry: Merging Biological and Mathematical Perspectives to Develop Modelling Competences

Reprinted from: *Educ. Sci.* **2021**, *11*, 611, doi:10.3390/educsci11100611 163

Amy M. Masnick and Bradley J. Morris

A Model of Scientific Data Reasoning

Reprinted from: *Educ. Sci.* **2022**, *12*, 71, doi:10.3390/educsci12020071 179

Sabine Meister and Annette Upmeier zu Belzen Analysis of Data-Based Scientific Reasoning from a Product-Based and a Process-Based Perspective Reprinted from: <i>Educ. Sci.</i> 2021 , <i>11</i> , 639, doi:10.3390/educsci11100639	199
Jennifer Schellinger, Patrick J. Enderle, Kari Roberts, Sam Skrob-Martin, Danielle Rhemer and Sherry A. Southerland Describing the Development of the Assessment of Biological Reasoning (ABR) Reprinted from: <i>Educ. Sci.</i> 2021 , <i>11</i> , 669, doi:10.3390/educsci11110669	221
Marvin Rost and Tarja Knuuttila Models as Epistemic Artifacts for Scientific Reasoning in Science Education Research Reprinted from: <i>Educ. Sci.</i> 2022 , <i>12</i> , 276, doi:10.3390/educsci12040276	241
Annette Upmeier zu Belzen, Paul Engelschalt and Dirk Krüger Modeling as Scientific Reasoning—The Role of Abductive Reasoning for Modeling Competence Reprinted from: <i>Educ. Sci.</i> 2021 , <i>11</i> , 495, doi:10.3390/educsci11090495	261
Liwei Wei, Carla M. Firetto, Rebekah F. Duke, Jeffrey A. Greene and P. Karen Murphy High School Students’ Epistemic Cognition and Argumentation Practices during Small-Group Quality Talk Discussions in Science Reprinted from: <i>Educ. Sci.</i> 2021 , <i>11</i> , 616, doi:10.3390/educsci11100616	273

Preface to “Scientific Reasoning in Science Education: From Global Measures to Fine-Grained Descriptions of Students’ Competencies”

In modern science- and technology-based societies, competencies that enable citizens to reason scientifically play a key role not only in science and technology-based careers but also for democratic co-determination (e.g., OECD, 2019). Developing these competencies is, hence, considered an important goal for science education in many countries around the globe (e.g., KMK, 2020; NRC, 2012).

Scientific reasoning competencies are defined as a complex construct that encompasses abilities such as identifying scientific problems, developing questions and hypotheses, categorizing and classifying entities, engaging in probabilistic reasoning, generating evidence through modeling, experimentation, etc., and communicating, evaluating, and scrutinizing claims (Lawson, 2004; NRC, 2012). These abilities require different forms of knowledge, such as content knowledge about the concepts of science, procedural knowledge about scientific methods, and epistemic knowledge of how such procedures warrant the claims that scientists advance (Osborne, 2014).

The research on scientific reasoning competencies is quite diverse. This diversity is—at least in part—caused by the manifold abilities that models of scientific reasoning comprise and the wide range of content, procedural, and epistemic knowledge that is deemed necessary to exercise these abilities. Differences exist, for instance, in the specific abilities that are addressed (e.g., applying the control-of-variables strategy: Reith & Nehring, 2020; handling of anomalous data: Chinn & Brewer, 2001; formulating questions and hypotheses: Vorholzer et al., 2016; developing and using models: Göhner & Krell, 2020). In addition, even studies that focus on similar abilities may use different theoretical frameworks and address different procedural and epistemic concepts (Vorholzer et al., 2016). Moreover, studies focus on a broad spectrum of respondents ranging from K-12 students (e.g., Koerber & Osterhaus, 2019; Mayer et al., 2014; Nehring et al., 2015; Vorholzer et al., 2016) to pre-service (e.g., Khan & Krell, 2019) and in-service teachers (e.g., Krell & Krüger, 2016).

Empirical research that focuses on scientific reasoning competencies typically describes the addressed competencies in a rather large-grained way. On a conceptual level, most studies offer a clear description of the addressed competencies, while the specific abilities, as well as the corresponding procedural and epistemic knowledge, are often less precisely defined (Vorholzer et al., 2016). For instance, a study may report that it focuses on students’ competencies to develop scientific investigations without stating whether that entails just knowledge of the control-of-variables strategy or also knowledge of strategies such as repeating measurements or measuring with large quantities. In addition, empirical studies often report aggregated measures, for instance, in the form of a global scientific reasoning competency measure or a global measure of their epistemic understanding (e.g., naïve vs. sophisticated) without stating what exactly students are (or are not) able to do or how they understand concepts related to scientific reasoning. It is important to note that the grain-size outlined above is completely sufficient when the goal of a study is, for instance, to investigate the effectiveness of a specific instructional intervention or to analyze the dimensionality of a competency model. Studies that utilize this grain-size have provided many vital insights regarding the modeling, assessment, and ways of fostering scientific reasoning competencies. However, we argue that more fine-grained perspectives have substantial benefits for instructional practice and research. For instance, detailed insights into students’ procedural and epistemic knowledge related to scientific

reasoning can inform teachers in designing instructions that match students' current understanding and specific learning needs. Such insights also provide manifold opportunities for further research, for instance, regarding the development of students' scientific reasoning competencies and the corresponding learning processes.

This book compiles empirical and theoretical contributions that seek to provide a more fine-grained perspective on scientific reasoning competencies, for instance, by providing precise descriptions of specific abilities and corresponding knowledge or by offering insights into the extent to which students of different age groups are able to reason scientifically. The contributions demonstrate the variety of conceptualizations of scientific reasoning in science education. Several contributions have based their research on well-established conceptualizations, such as formulating research questions, generating hypotheses, planning experiments, observing and measuring, preparing data for analysis, and drawing conclusions (e.g., Bicak et al.). Others have broadened their scope and discuss aspects that are somewhat "on the sidelines" of what is typically considered scientific reasoning, such as the relevance of conceptual knowledge for reasoning in a specific context (Schellinger et al.) and through reasoning on controversial science issues (Beniermann et al.). Most of the contributions address the abilities related to experimentation and modeling (e.g., Khan & Krell; Upmeier zu Belzen et al.).

The contributions in the book are ordered by their conceptualizations of scientific reasoning. After the editorial, six contributions follow, which address scientific reasoning in general (Bicak et al.; Hilfert-Rüppel et al.; Khan & Krell; Krell et al.; Mahler et al.; Schlatter et al.). The second part of the book comprises nine contributions, which address specific aspects of scientific reasoning such as modeling- or data-based reasoning (Beniermann et al.; Cabello et al.; Lang et al.; Masnick & Morris; Meister & Upmeier zu Belzen; Schellinger et al.; Rost & Knuuttila; Upmeier zu Belzen et al.; Wei et al.).

References:

- Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction, 19*, 323–393.
- Göhner, M. & Krell, M. (2020). Preservice science teachers' strategies in scientific reasoning: The case of modeling. *Research in Science Education, 52*, 395–414.
- Khan, S. & Krell, M. (2019). Scientific reasoning competencies: A case of preservice teacher education. *Canadian Journal of Science, Mathematics and Technology Education, 19*, 446–464.
- KMK [Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der BRD] (Eds.). (2020). *Bildungsstandards im Fach Biologie für die Allgemeine Hochschulreife*. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2020/2020_06_18-BildungsstandardsAHR.Biologie.pdf
- Koerber, S. & Osterhaus, C. (2019). Individual differences in early scientific thinking: Assessment, cognitive influences, and their relevance for science learning. *Journal of Cognition and Development, 20*, 510–533.
- Krell, M., & Krüger, D. (2016). Testing models: A key aspect to promote teaching activities related to models and modelling in biology lessons? *Journal of Biological Education, 50*, 160–173.
- Lawson, A. (2004). The nature and development of scientific reasoning: A synthetic view. *International Journal of Science and Mathematics Education, 2*, 307–338.
- Mayer, D., Sodian, B., Koerber, S. & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning and Instruction, 29*, 43–55.
- Nehring, A., Nowak, K. H., Upmeier zu Belzen, A. & Tiemann, R. (2015). Predicting

students' skills in the context of scientific inquiry with cognitive, motivational, and sociodemographic variables. *International Journal of Science Education*, 37, 1343–1363.

NRC [National Research Council] (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

OECD (2019). *Conceptual learning framework: Learning Compass 2030*. https://www.oecd.org/education/2030-project/teaching-and-learning/learning/learning-compass-2030/OECD_Learning_Compass_2030_concept_note.pdf

Osborne, J. (2014). Scientific practices and inquiry in the science classroom. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research on science education*. Volume 2 (pp. 579–599). New York: Routledge/Taylor & Francis Group.

Reith, M., & Nehring, A. (2020). Scientific reasoning and views on the nature of scientific inquiry: testing a new framework to understand and model epistemic cognition in science. *International Journal of Science Education*, 42, 2716–2741.

Vorholzer, A., von Aufschnaiter, C., & Kirschner, S. (2016). Entwicklung und Erprobung eines Tests zur Erfassung des Verständnisses experimenteller Denk- und Arbeitsweisen. *Zeitschrift für Didaktik der Naturwissenschaften*, 22, 25–41.

Moritz Krell, Andreas Vorholzer, and Andreas Nehring
Editors

Editorial

Scientific Reasoning in Science Education: From Global Measures to Fine-Grained Descriptions of Students' Competencies

Moritz Krell ^{1,*}, Andreas Vorholzer ² and Andreas Nehring ³

¹ Department Biology Education, IPN—Leibniz Institute for Science and Mathematics Education, Olshausenstr. 62, 24118 Kiel, Germany

² School of Social Sciences and Technology, Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany; andreas.vorholzer@tum.de

³ Institute for Science Education, Leibniz Universität Hannover, Am Kleinen Felde 30, 30167 Hannover, Germany; nehring@idn.uni-hannover.de

* Correspondence: krell@leibniz-ipn.de

Citation: Krell, M.; Vorholzer, A.; Nehring, A. Scientific Reasoning in Science Education: From Global Measures to Fine-Grained Descriptions of Students' Competencies. *Educ. Sci.* **2022**, *12*, 97. <https://doi.org/10.3390/educsci12020097>

Received: 24 January 2022

Accepted: 26 January 2022

Published: 30 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In modern science- and technology-based societies, competencies that allow citizens to reason scientifically play a key role for science- and technology-based careers as well as for democratic co-determination (e.g., [1]). Most challenges that societies are facing (e.g., climate change, food and energy supply, or—recently—the COVID-19 pandemic) and also many questions that we encounter as individuals in our everyday lives (e.g., vaccination, nutrition, electric mobility) are strongly related to science. Therefore, participating in the public discourse on societal challenges as well as making informed decisions in one's own life requires not only a basic understanding of science-based concepts, but also an understanding of the ways in which scientists think and reason. Developing scientific reasoning competencies is, hence, considered an important goal of science education in many countries around the globe (e.g., [2,3]).

The terms “scientific reasoning” and “scientific reasoning competencies” are not consistently defined in the literature. In this editorial, we will, therefore, first analyze existing definitions of scientific reasoning in science education and outline similarities and differences between them. On the basis of this analysis, we will argue that it is important to adopt a fine-grained perspective that accounts for the specific abilities, as well as the corresponding knowledge, usually included under the umbrella term “scientific reasoning”. Second, we will illustrate the potential benefits of conceptualizing scientific reasoning as a competency rather than merely as abilities and/or knowledge. Third, we will use these theoretical considerations to provide a structured overview of the contributions in this Special Issue.

2. Definitions of Scientific Reasoning

Scientific reasoning is typically conceptualized as a complex construct that encompasses a wide range of abilities as well as corresponding knowledge; clear definitions, however, have only seldom been proposed. On the basis of the notion of competency (see below), Mathesius et al. [4] defined scientific reasoning as the “competencies that are needed to understand the processes through which scientific knowledge is acquired” (p. 94). It is important to note that this definition of scientific reasoning has a significant overlap with the related goals of science education, such as scientific inquiry or scientific thinking. This overlap can be seen in the fact that these terms often refer to similar abilities and knowledge. These abilities, which are typically summarized under the term “scientific reasoning”, comprise the following: identifying scientific problems, developing questions and hypotheses, categorizing and classifying entities, engaging in probabilistic

reasoning, generating evidence through modeling or experimentation, and communicating, evaluating, and scrutinizing claims (e.g., [5–8]). The corresponding knowledge—that is, knowledge necessary to reason scientifically in a given context—comprises content knowledge, procedural knowledge, and epistemic knowledge. *Content knowledge* about science concepts (e.g., theories, laws, definitions) refers to knowledge about the objects that science reasons with and about [9]. This kind of knowledge is necessary as a basis for scientific reasoning in a specific context (“knowing what”; [8]). For instance, to develop meaningful scientific hypotheses about the variables that impact the efficiency of a solar panel, students need a sufficient conceptual understanding of the variables that are potentially relevant in that context (e.g., voltage, current, electric energy, inclination, charge separation). *Procedural knowledge* refers to knowledge about the rules, practices, and strategies that the processes of scientific reasoning are based upon (“knowing how”; [8]). For instance, in the example given above, students require not only knowledge about science concepts but also knowledge of the rules on how to formulate a suitable scientific hypothesis (e.g., hypotheses should be as precise as possible; hypotheses have to be falsifiable; see [10–12]). In addition to content and procedural knowledge, students should also have an epistemic understanding about why scientific reasoning (or a specific part of it, e.g., developing a scientific hypothesis) is important and how it contributes to building reliable scientific knowledge (“knowing why”; [8,12]). Such *epistemic knowledge* is not directly necessary to develop a well-formulated and falsifiable scientific hypothesis, but it is mandatory to ensure that formulating hypotheses is more than rote performance [8,13].

Early and seminal descriptions of scientific reasoning (or similar constructs such as logical thinking or scientific thinking) already appeared in the work of the developmental psychologists Inhelder and Piaget [14] about the stages of human thinking. In their work, evaluating claims based on observations was, for instance, part of the highest cognitive stage (formal operational reasoning). This implied, however, one single cognitive ability and not a multidimensional nature of reasoning, as stressed by later approaches.

In the late 1980s, Klahr and colleagues proposed the very influential model of scientific discovery as dual search (SDDS) as a way to capture scientific reasoning (e.g., [15,16]). In the SDDS model, scientific reasoning is conceptualized as a search in the following two problem spaces: the hypothesis space and the experiment space. On the basis of this “dual search”, Klahr and colleagues distinguished between the scientific reasoning processes of hypotheses generation (“search hypothesis”), experimental design (“test hypothesis”), and hypotheses evaluation (“evaluate evidence”; [16], p. 33). In the SDDS model, scientific reasoning is positioned within the context of problem solving (which, in retrospect, can be envisioned as a link to the notion of competency; see below). Additionally, the idea of a single cognitive ability (see [14]) is extended to a differentiation between multiple distinct abilities. The SDDS model was adopted in a number of different studies with backgrounds in psychology and science education (e.g., [17–19]). Furthermore, the three-phase structure is still a prominent way of modeling scientific reasoning in science education [17].

More recent approaches have shown a diversity of conceptualizations, particularly by either broadening what is considered to be scientific reasoning *beyond* the realm of experimental hypothesis testing (aligned horizontally in Figure 1) or differentiating between more or different sub-processes of scientific reasoning *within* that realm (aligned vertically in Figure 1). The former can be observed, for instance, in the work of Cullinane, Erduran, and Wooding [20], who not only investigated experimental hypothesis testing (manipulative types of methods) but also at non-manipulative approaches. Another example is the work of Kind and Osborne [9], who proposed six different styles of reasoning (mathematical deduction, experimental evaluation, hypothetical modeling, categorization and classification, probabilistic reasoning, and historical-based evolutionary reasoning). Here, the SDDS model and related descriptions of scientific reasoning would probably just represent a more detailed description of the experimental evaluation style, while all the other styles go beyond what is typically understood as scientific reasoning. An example of the latter type of conceptualization is the work by Krüger et al. [21], who distinguished between

four process-related dimensions (called “sub-competencies”) of conducting scientific investigations by formulating questions, generating hypotheses, planning investigations, and analyzing data and drawing conclusions. White and Frederiksen [22] even went a step further by distinguishing between the six sub-processes of questioning, hypothesizing, investigating, analyzing, modeling, and evaluating.

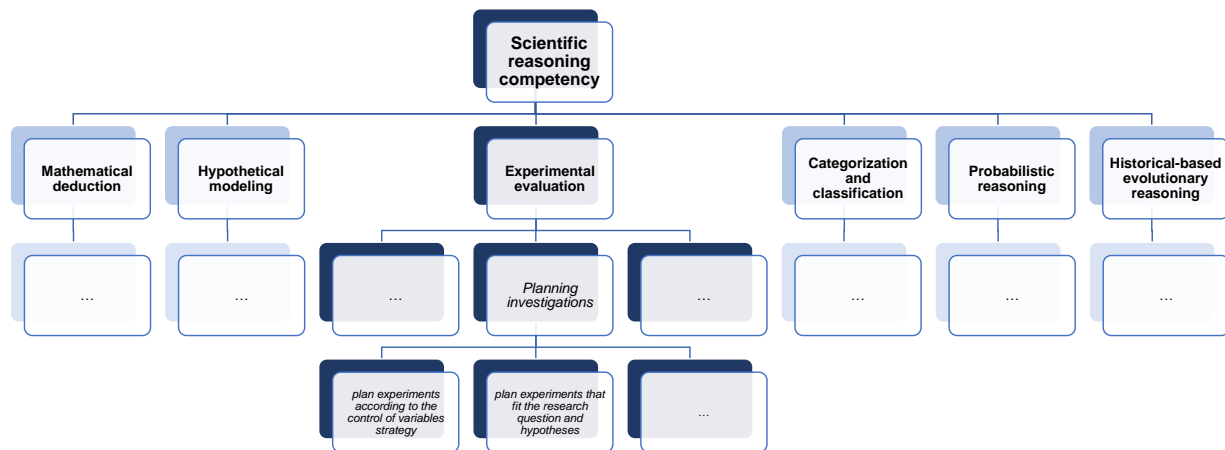


Figure 1. Spectrum of (sub-)competencies typically associated with scientific reasoning.

In sum, most of the more recent conceptualizations of scientific reasoning competencies share the assumption that there are several sub-processes that can be modeled as separate dimensions of scientific reasoning (Figure 1). However, as outlined above, the ways in which these dimensions are shaped vary significantly. To capture and systemize the (increasing) variance in conceptualizations of (and research on) scientific reasoning (competencies), three different aspects have been identified, in which conceptualizations of scientific reasoning competencies may differ: “(a) the skills they include, (b) whether there is a general, uniform scientific reasoning ability or, rather, more differentiated dimensions of scientific reasoning, and (c) whether they assume scientific reasoning to be domain-general or domain-specific” [23] (p. 80). While these aspects are certainly very helpful in capturing the variance in conceptualizations of scientific reasoning, we argue that there is even more to be discovered beneath the surface.

3. Grain Sizes: A Source of Relevant but Seldom Addressed Variance

Empirical research that focuses on scientific reasoning typically describes the addressed competency as well as the corresponding outcomes with a rather large “grain size”. For instance, most studies offer a clear conceptual description of the addressed competency (the larger grain), while the specific abilities as well as the corresponding procedural and epistemic knowledge (i.e., the smaller grains that “make up” these competencies) often remain unclear (see discussion in [24]). Therefore, different conceptualizations of scientific reasoning may seem similar regarding the competency/sub-competencies they comprise, but may still address different abilities and/or the different content, procedural, and epistemic knowledge associated with them [24]. For example, the competency of “planning scientific investigations” may include the ability to (a) “plan investigations according to the control of variables strategy”, (b) “plan investigations considering repeated measurements”, (c) “plan investigations that fit the research question and hypotheses”, or a combination of these abilities (Figure 1). All of the aforementioned abilities clearly belong to the competency of “planning scientific investigations”; however, the abilities addressed in an assessment or an instructional intervention have considerable consequences for the tasks students are working on and the knowledge they need to solve them, and, hence, may significantly impact the results [24]. For instance, (a) requires procedural knowledge of the control of variables strategy as well as of terminology such as the dependent variable,

independent variable, etc., and epistemic knowledge about the function of controlling variables. In contrast, (b) requires procedural knowledge of strategies for how to determine the appropriate number of repetitions of measurement procedures as well as epistemic knowledge about why repeating measurements is important.

This example illustrates that the results of studies that have adopted one specific conceptualization of a scientific reasoning competency (i.e., the larger grain) may vary considerably depending on which abilities and corresponding knowledge (i.e., smaller grains) are considered to be part of that specific conceptualization. Differences on a fine-grained level between the conceptualizations of scientific reasoning adopted in various studies are not per se a problem. Given that the construct of scientific reasoning is manifold, it is often even necessary to focus on a few specific abilities and the corresponding knowledge. However, without a fine-grained description of the chosen conceptualization, these differences cannot be properly accounted for. This is also emphasized by Shavelson [25], who argues that a clear construct definition is necessary in order to develop authentic assessments that are capable of holistically assessing all (or as many as possible) of the components of a given competency (holistic approach of competency assessment). From a more analytical point of view, it has been proposed that all skills of a given competency need to be clearly defined so that tasks (i.e., indicators) can be developed for each [26].

Aside from improving the comparability between studies, a more fine-grained perspective may have additional positive effects, for instance, regarding the communication and application of empirical findings. Empirical studies often report aggregated measures (larger grains), for example, in the form of a single (i.e., global) measure of students' scientific reasoning competency, their procedural understanding of the control of variables strategy, or their epistemic understanding (e.g., naïve vs. sophisticated). This grain size is sufficient when the goal of a study is, for example, to investigate the effectiveness of a specific instructional intervention or to obtain system-monitoring data. However, goals such as designing instruction that matches students' current understanding and specific learning needs require more detailed insights. From an educational point of view, insights into which procedural and epistemic knowledge students have and which they lack and into which abilities they have mastered and which they have not mastered yet is of vital importance for designing instructional interventions.

We argue that research would profit from examining and reporting scientific reasoning by using more fine-grained perspectives in addition to the rather global grain sizes typically reported. Linking these fine-grained and global perspectives would make it possible to compare and even to coordinate research approaches. Additionally, research on the development of students' scientific reasoning competencies and the corresponding learning processes—a key challenge for science education research on scientific reasoning competencies—would benefit from linking fine-grained and global perspectives and not only sticking to the one or the other perspective.

4. What the Notion of Competency Implies for Research on Scientific Reasoning in Science Education

Following the increasing attention that the notion of describing learning outcomes as competencies has received in recent decades (in particular, in German-speaking countries), an increasing number of studies that (explicitly or implicitly) describe scientific reasoning as a competency have emerged (e.g., [7,27]). Competencies can be defined broadly as “dispositions that are acquired by learning and needed to successfully cope with certain situations or tasks in specific domains” [28] (p. 9). Most conceptualizations of competencies in science education share a common core of characteristic features (see [29–31]): First, competencies comprise cognitive as well as motivational, volitional, ethical, social, and further dispositions [32]. Second, these dispositions are domain-specific and learnable [28], and, third, these dispositions are transferable in the sense that they enable individuals to perform successfully within the same family of problems [33]. More recent definitions of the term “competency” have also emphasized the role of situation-specific skills such as

“perception” or “decision-making”, which mediate the transition of dispositions to actual performance [34]. The notion of competency and its development adds new and interesting facets to research on scientific reasoning, which, as we will detail below, are yet to be explored.

In science education research, scientific reasoning is often conceptualized in terms of skills or abilities (e.g., [35,36]) that require specific knowledge, and corresponding studies typically focus on learners’ cognitive dispositions. A similar focus on cognitive dispositions can also be observed in studies that describe scientific reasoning as a competency (e.g., [37]). There is no doubt that cognitive dispositions are an important element of scientific reasoning competency. However, given that a key goal of promoting scientific reasoning is to allow (future) citizens to reason scientifically in their personal and professional lives (e.g., [1]), we argue that it is important to consider all elements of the notion of competency outlined above in conceptualizations of scientific reasoning competency (e.g., cognitive, motivational, and social). For instance, to use scientific reasoning to make an informed decision on a science- or technology-related topic or problem in an everyday situation, it is not sufficient to be *able* to reason scientifically (i.e., to have specific abilities and corresponding knowledge). One also has to realize that the particular situation actually requires scientific reasoning or even a specific style of scientific reasoning (perception) and, at the same time, one needs to be *willing* to apply the related abilities and knowledge (motivational dispositions). Therefore, we argue that considering scientific reasoning as a competency has multiple fruitful implications for future research: First, it suggests that we require the means to assess and promote learners’ cognitive dispositions (e.g., abilities and knowledge related to scientific reasoning) but also means to assess and promote their motivational dispositions (e.g., willingness to engage in scientific reasoning, beliefs about the usefulness of scientific reasoning). Second, alongside focusing on students’ cognitive, motivational, and social dispositions, it is also important to identify the situation-specific skills that facilitate or hinder the process by which these dispositions are translated into performance as well as to find the means to assess and foster them. Lastly, and probably most challengingly, it is often assumed that promoting students’ abilities to plan and scrutinize scientific investigations or to analyze and interpret data (i.e., fostering their dispositions and/or situation-specific skills) helps them to solve scientific problems as well as to make informed decisions in their everyday lives (i.e., improves their performance); the relationship between dispositions, situation-specific skills, and performance in a (close to) real-life situation, however, has—to the best of our knowledge—only rarely been investigated so far (examples in [38]). In sum, considering scientific reasoning as a competency is more than a rebranding; it emphasizes the role of motivational and social dispositions, situation-specific skills, and actual performance and, thereby, outlines a promising agenda for science education research.

5. Contributions in This Special Issue

The contributions in this Special Issue demonstrate the variety of conceptualizations of scientific reasoning in science education. Several contributions have based their research on well-established conceptualizations, such as formulating research questions, generating hypotheses, planning experiments, observing and measuring, preparing data for analysis, and drawing conclusions (e.g., [39]). Others have broadened their scope and discuss aspects that are somewhat “on the sidelines” of what is typically considered scientific reasoning, such as the relevance of conceptual knowledge for reasoning in a specific context [40] and through reasoning on controversial science issues [41]. Most of the contributions address the abilities related to experimentation and modeling (e.g., [39,42,43]); this mirrors the focus of science education on these two styles of reasoning. However, this focus has been criticized as “an impoverished account” of scientific reasoning [9] (p. 17).

The articles in this Special Issue also show how scientific reasoning competencies can be addressed based on smaller and larger grain sizes. For example, Bicak et al. [39] conceptualize scientific reasoning in a six-step approach (see above) based on a larger grain size. In their work, the authors provide an insight into how they operationalized

scientific reasoning in a coding manual for video recordings and written records in a hands-on task. For the sub-competency of observing and measuring, for instance, students reached the highest proficiency level when their observations or measurements were purposeful, exhaustive, and correct and their data was recorded correctly by using a suitable method of measurement. The approach of coding outlined in [39] indicates that scientific correctness is at the center of this conceptualization and not measurement repetition or error analysis, as it might have been in a physics approach. Another example is provided by Khan and Krell [42], who describe a two-dimensional competency-based approach to scientific reasoning (conducting scientific investigations and using scientific models) and also define sub-competencies of scientific reasoning and the associated abilities, including the necessary procedural and epistemic knowledge (based on [11]). They argue that the ability of generating hypotheses requires the knowledge that hypotheses are empirically testable, intersubjectively comprehensible, clear, logically consistent, and compatible with an underlying theory [42] (p. 3). This also presents the opportunity to link global and fine-grained perspectives on scientific reasoning competencies.

Considering the different aspects of competency outlined above, it is evident that the contributions in this Special Issue focus primarily on cognitive dispositions related to scientific reasoning competencies. However, there are interesting exceptions. For instance, the study of Beniermann, Mecklenburg, and Upmeier zu Belzen [41] investigated reasoning processes in the context of controversial science issues. They argue that decisions or attitudes regarding these issues depend “highly on individual norms and values” [41] (p. 2) and are, therefore, not only dependent on learners’ content knowledge or their ability to reason scientifically. The authors were able to identify not only intersubjective but also subjective types of reasoning in their analyses; this finding can be interpreted as empirical support for the assumed role of non-cognitive dispositions in scientific reasoning processes. Furthermore, their work suggests that controversial science issues might be a promising context in which to further investigate the non-cognitive components of scientific reasoning competency. Another interesting example is presented by Meister and Upmeier zu Belzen [44], who investigated the role of anomalous data in scientific reasoning processes. Their approach highlights that learners’ perceived relevance and subsequent appraisal of data depend on their prior conceptions. As learners’ prior conceptions are presumably very situation-specific, perception and appraisal processes may serve as an interesting example of the situation-specific skills that mediate the transformation of dispositions into performance [34]. In a similar vein, the authors were also able to show that the role of the perceived acceptance of an interpretation by others is—in some cases—more relevant than evidential considerations, which, again, illustrates the role of situation-specific skills and emotional dispositions in scientific reasoning processes. In sum, among the contributions in this Special Issue, there are some interesting approaches that consider both cognitive dispositions and motivational dispositions or situation-specific skills. However, the role that aspects of competency other than cognitive dispositions play in scientific reasoning remains a rather underrepresented field of research that, in our opinion, provides considerable potential for further research.

Author Contributions: Conceptualization: M.K., A.N. and A.V.; Writing—Original Draft Preparation: M.K., A.N. and A.V.; Writing—Review and Editing: M.K., A.N. and A.V.; Visualization: M.K.; Supervision: M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. OECD. *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics, Financial Literacy and Collaborative Problem Solving (Revised Edition)*; OECD Publishing: Paris, France, 2017.
2. KMK. *Bildungsstandards im Fach Biologie für die Allgemeine Hochschulreife [Educational Standards in Biology for the Higher Education Entrance Qualification]*; Wolters Kluwer: Hürth, Germany, 2020.

3. NRC. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*; National Academies Press: Washington, DC, USA, 2012.
4. Mathesius, S.; Hartmann, S.; Upmeier zu Belzen, A.; Krüger, D. Scientific reasoning as an aspect of pre-service biology teacher education. In *The Future of Biology Education Research. Proceedings of the 10th conference of European Researchers in Didactics of Biology (ERIDOB)*; Tal, T., Yarden, A., Eds.; Technion: Haifa, Israel, 2016; pp. 93–110.
5. Hartmann, S.; Upmeier zu Belzen, A.; Krüger, D.; Pant, H.A. Scientific reasoning in higher education. *Z. Psychol.* **2015**, *223*, 47–53. [CrossRef]
6. Lawson, A.E. The nature and development of scientific reasoning. *A synthetic view. Int. J. Sci. Math. Educ.* **2004**, *2*, 307–338. [CrossRef]
7. Osborne, J. Teaching scientific practices: Meeting the challenge of change. *J. Sci. Teach. Educ.* **2014**, *25*, 177–196. [CrossRef]
8. Osborne, J. Scientific practices and inquiry in the science classroom. In *Handbook of Research on Science Education*; Lederman, N., Abell, S., Eds.; Routledge: New York, NY, USA, 2014; pp. 579–599.
9. Kind, P.M.; Osborne, J. Styles of scientific reasoning. A cultural rationale for science education? *Sci. Educ.* **2017**, *101*, 8–31. [CrossRef]
10. Arnold, J.C.; Mühling, A.; Kremer, K. Exploring core ideas of procedural understanding in scientific inquiry using educational data mining. *Res. Sci. Technol. Educ.* **2021**, *70*, 1–21. [CrossRef]
11. Mathesius, S.; Upmeier zu Belzen, A.; Krüger, D. Kompetenzen von Biologiestudierenden im Bereich der naturwissenschaftlichen Erkenntnisgewinnung: Entwicklung eines Testinstruments [Competencies of biology students in the field of scientific inquiry: Development of a testing instrument]. *Erkenn. Biol.* **2014**, *13*, 73–88.
12. Vorholzer, A.; von Aufschnaiter, C.; Boone, W.J. Fostering upper secondary students' ability to engage in practices of scientific investigation: A comparative analysis of an explicit and an implicit instructional approach. *Res. Sci. Educ.* **2020**, *50*, 333–359. [CrossRef]
13. Berland, L.K.; Schwarz, C.V.; Krist, C.; Kenyon, L.; Lo, A.S.; Reiser, B.J. Epistemologies in practice. Making scientific practices meaningful for students. *J. Res. Sci. Teach.* **2016**, *53*, 1082–1112. [CrossRef]
14. Inhelder, B.; Piaget, J. *The Growth of Logical Thinking from Childhood to Adolescence: An Essay on the Construction of Formal Operational Structures*; Routledge & Kegan Paul: London, UK, 1958.
15. Klahr, D.; Carver, S.M. Scientific thinking about scientific thinking. *Monogr. Soc. Res. Child Dev.* **1995**, *60*, 137–151. [CrossRef]
16. Klahr, D.; Dunbar, K. Dual space search during scientific reasoning. *Cogn. Sci.* **1988**, *12*, 1–48. [CrossRef]
17. Emden, M.; Sumfleth, E. Assessing students' experimentation processes in guided inquiry. *Int. J. Sci. Math. Educ.* **2016**, *14*, 29–54. [CrossRef]
18. Hammann, M.; Phan, T.; Ehmer, M.; Grimm, T. Assessing pupils' skills in experimentation. *J. Biol. Educ.* **2008**, *42*, 66–72. [CrossRef]
19. Krell, M. Schwierigkeitserzeugende Aufgabenmerkmale bei Multiple-Choice-Aufgaben zur Experimentierkompetenz im Biologieunterricht: Eine Replikationsstudie [Difficulty-generating task characteristics of multiple-choice-tasks for assessing experimental competencies: A replication study]. *Z. Didakt. Nat.* **2018**, *24*, 1–15.
20. Cullinane, A.; Erduran, S.; Wooding, S.J. Investigating the diversity of scientific methods in high-stakes chemistry examinations in England. *Int. J. Sci. Educ.* **2019**, *41*, 2201–2217. [CrossRef] [PubMed]
21. Krüger, D.; Hartmann, S.; Nordmeier, V.; Upmeier zu Belzen, A. Measuring scientific reasoning competencies. In *Student Learning in German Higher Education*; Zlatkin-Troitschanskaia, O., Pant, H., Toepper, M., Lautenbach, C., Eds.; Springer: Wiesbaden, Germany, 2020; pp. 261–280.
22. White, B.; Frederiksen, J. A theoretical framework and approach for fostering metacognitive development. *Educ. Psychol.* **2005**, *40*, 211–223. [CrossRef]
23. Opitz, A.; Heene, M.; Fischer, F. Measuring scientific reasoning. A review of test instruments. *Educ. Res. Eval.* **2017**, *23*, 78–101. [CrossRef]
24. Vorholzer, A.; von Aufschnaiter, C.; Kirschner, S. Entwicklung und Erprobung eines Tests zur Erfassung des Verständnisses experimenteller Denk- und Arbeitsweisen [Development of an instrument to assess students' knowledge of scientific inquiry]. *Z. Didakt. Nat.* **2016**, *22*, 25–41.
25. Shavelson, R.J. On an approach to testing and modeling competence. *Educ. Psychol.* **2013**, *48*, 73–86. [CrossRef]
26. Frey, A. Strukturierung und Methoden zur Erfassung von Kompetenz. *Bild. Erzieh.* **2006**, *59*, 125–166. [CrossRef]
27. Upmeier zu Belzen, A.; van Driel, J.H.; Krüger, D. Introducing a framework for modeling competence. In *Towards a Competence-Based View on Models and Modeling in Science Education*; Upmeier zu Belzen, A., Krüger, D., van Driel, J.H., Eds.; Springer: Cham, Switzerland, 2019; pp. 3–19.
28. Klieme, E.; Hartig, J.; Rauch, D. The concept of competence in educational contexts. In *Assessment of Competencies in Educational Contexts*; Hartig, J., Klieme, E., Leutner, D., Eds.; Hogrefe: Göttingen, Germany, 2008; pp. 3–22.
29. Krell, M. Vorstellung und Kompetenz: Vergleich und Vorschlag einer Zusammenführung zweier zentraler Konzepte der naturwissenschaftsdidaktischen Forschung [Conception and competency: Comparison and proposal of a synthesis of two central concepts of science education research]. In *Vorstellungsforschung in der Biologiedidaktik: Theorie, Kompetenz, Diagnose, Intervention*; Reinisch, B., Helbig, K., Krüger, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2020; pp. 69–82.

30. Reith, M.; Nehring, A. Scientific reasoning and views on the nature of scientific inquiry: Testing a new framework to understand and model epistemic cognition in science. *Int. J. Sci. Educ.* **2020**, *42*, 2716–2741. [CrossRef]
31. Vorholzer, A.; von Aufschnaiter, C. Dimensionen und Ausprägungen fachinhaltlicher Kompetenz in den Naturwissenschaften: Ein Systematisierungsversuch [Dimensions and levels of subject-matter competency—An attempt to systematize research in science]. *Z. Didakt. Nat.* **2020**, *26*, 1–18.
32. Weinert, F.E. Concept of competence: A conceptual clarification. In *Defining and Selecting Key Competencies*; Rychen, D.S., Salganik, L.H., Eds.; Hogrefe: Kirkland, WA, USA, 2001; pp. 45–65.
33. Rychen, D.; Salganik, L. A holistic model of competence. In *Key Competencies for a Successful Life and a Well-Functioning Society*; Rychen, D., Salganik, L., Eds.; Hogrefe & Huber: Cambridge, UK, 2003; pp. 41–62.
34. Blömeke, S.; Gustafsson, J.-E.; Shavelson, R.J. Beyond dichotomies. *Z. Psychol.* **2015**, *223*, 3–13. [CrossRef]
35. Nehring, A.; Nowak, K.H.; zu Belzen, A.U.; Tiemann, R. Predicting students' skills in the context of scientific inquiry with cognitive, motivational, and sociodemographic variables. *Int. J. Sci. Educ.* **2015**, *37*, 1343–1363. [CrossRef]
36. Van der Graaf, J.; van de Sande, E.; Gijssels, M.; Segers, E. A combined approach to strengthen children's scientific thinking: Direct instruction on scientific reasoning and training of teacher's verbal support. *Int. J. Sci. Educ.* **2019**, *41*, 1119–1138. [CrossRef]
37. Stiller, J.; Hartmann, S.; Mathesius, S.; Straube, P.; Tiemann, R.; Nordmeier, V.; Krüger, D.; Upmeyer zu Belzen, A. Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assess. Eval. High. Educ.* **2016**, *41*, 721–732. [CrossRef]
38. Reith, M.; Nehring, A. Experimentieren in der „Kompetenztrias“ aus Disposition, Prozess und Produkt erfassen, modellieren und fördern. [Assessing, modeling, and fostering experimenting using the competency-triad—Disposition, process and product]. In *Naturwissenschaftlicher Unterricht und Lehrerbildung im Umbruch?* Habig, S., Ed.; Universität Duisburg-Essen: Duisburg, Germany, 2021; pp. 537–540.
39. Bıcak, B.E.; Borchert, C.E.; Höner, K. Measuring and fostering preservice chemistry teachers' scientific reasoning competency. *Educ. Sci.* **2021**, *11*, 496. [CrossRef]
40. Schellinger, J.; Enderle, P.J.; Roberts, K.; Skrob-Martin, S.; Rhemer, D.; Southerland, S.A. Describing the Development of the Assessment of Biological Reasoning (ABR). *Educ. Sci.* **2021**, *11*, 669. [CrossRef]
41. Beniermann, A.; Mecklenburg, L.; Upmeyer zu Belzen, A. Reasoning on controversial science issues in science education and science communication. *Educ. Sci.* **2021**, *11*, 522. [CrossRef]
42. Khan, S.; Krell, M. Patterns of Scientific Reasoning Skills among Pre-Service Science Teachers: A Latent Class Analysis. *Educ. Sci.* **2021**, *11*, 647. [CrossRef]
43. Upmeyer zu Belzen, A.; Engelschalt, P.; Krüger, D. Modeling as Scientific Reasoning: The Role of Abductive Reasoning for Modeling Competence. *Educ. Sci.* **2021**, *11*, 495. [CrossRef]
44. Meister, S.; Upmeyer zu Belzen, A. Analysis of Data-Based Scientific Reasoning from a Product-Based and a Process-Based Perspective. *Educ. Sci.* **2021**, *11*, 639. [CrossRef]

Article

Measuring and Fostering Preservice Chemistry Teachers' Scientific Reasoning Competency

Besim Enes Bicak *, Cornelia Eleonore Borchert  and Kerstin Höner *Institut für Fachdidaktik der Naturwissenschaften, Technische Universität Braunschweig,
38106 Braunschweig, Germany; cornelia.borchert@tu-braunschweig.de

* Correspondence: b.bicak@tu-braunschweig.de (B.E.B.); k.hoener@tu-braunschweig.de (K.H.)

Abstract: Developing scientific reasoning (SR) is a central goal of science-teacher education worldwide. On a fine-grained level, SR competency can be subdivided into at least six skills: *formulating research questions, generating hypotheses, planning experiments, observing and measuring, preparing data for analysis, and drawing conclusions*. In a study focusing on preservice chemistry teachers, an organic chemistry lab course was redesigned using problem-solving experiments and SR video lessons to foster SR skills. To evaluate the intervention, a self-assessment questionnaire was developed, and a performance-based instrument involving an experimental problem-solving task was adapted to the target group of undergraduates. The treatment was evaluated in a pre-post design with control group (cook-book experiments, no SR video lessons) and alternative treatment group (problem-solving experiments, unrelated video lessons). Interrater reliability was excellent (ρ from 0.915 to 1.000; ICC (A1)). Data analysis shows that the adapted instrument is suitable for university students. First insights from the pilot study indicate that the cook-book lab (control group) only fosters students' skill in *observing and measuring*, while both treatment groups show an increase in *generating hypotheses* and *planning experiments*. No pretest-posttest differences were found in self-assessed SR skills in the treatment groups. Instruments and data are presented and discussed.

Citation: Bicak, B.E.; Borchert, C.E.; Höner, K. Measuring and Fostering Preservice Chemistry Teachers' Scientific Reasoning Competency. *Educ. Sci.* **2021**, *11*, 496. <https://doi.org/10.3390/educsci11090496>

Keywords: scientific reasoning; scientific inquiry; science education; chemistry; teacher education; assessment

Academic Editors: Moritz Krell,
Andreas Vorholzer and
Andreas Nehring

Received: 23 July 2021

Accepted: 30 August 2021

Published: 3 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The study of scientific thinking dates back nearly a century and has been the interest of psychologists and science educators alike [1,2]. The origins lie at Piaget's [3] theory of stages of cognitive development, formulating that adolescents are able to evaluate evidence and build hypothetical terms. One main focus of research covers the development of domain-general strategies of reasoning and problem-solving [2]. Terminology varies respectively by discipline as well as focus on research or teaching, yielding terms such as scientific reasoning, critical thinking (with regard to a science context), scientific discovery, scientific inquiry, or inquiry learning [2,4]. For the purpose of this article, scientific inquiry is used to describe teaching methods and activities aimed at the process of gaining scientific knowledge [5,6], whereas scientific reasoning (SR) refers to the cognitive skills required during a scientific inquiry activity [1,6,7]. Due to its cognitive nature, SR is viewed to be a competency consisting of a complex set of skills [8,9].

SR is needed for acquiring scientific knowledge [10], making it an essential competency in science and one of the most important key competencies in societies, increasing their technical progress. Therefore, developing SR is a central goal of school science as well as science-teacher preparation worldwide [11–19]. In the context of teaching, SR can be fostered by inquiry teaching activities [20]. Teachers, however, perceive the integration of inquiry into their lessons to be a difficult task [21,22], particularly if they have never conducted inquiry experiments before [5]. SR and methods of inquiry teaching should therefore already be introduced in preservice science-teacher education [5,15,23].

1.1. Models for Scientific Reasoning Competency

SR can be studied in problem-solving situations, making it accessible to assessment [1,24]. Using a “simulated discovery context” [1] (p. 3) gives detailed insight into the underlying processes while maintaining control over situational boundaries and prior knowledge needed by the participants. So far, several models of scientific reasoning ranging from global measures to more fine-grained descriptions have been proposed, yielding various instruments for measuring SR competency [25]. Models can be differentiated by target group, i.e., school and/or university context or other target groups [25], and by their purposes, such as assessment [10], description of learners’ activities in the process of SR, and problem-solving [26,27] or the description of important aspects of inquiry teaching [28]. Following Klahr and Dunbar’s [29] “Dual Space Search,” SR models usually cover the domain “conducting scientific investigations” [24] and exhibit at least a three-fold structure involving the facets “hypothesis,” “experiment,” and “conclusion”. However, the respective models differ in the degree of sub-division into up to 9 facets (see Figure 1) [10,24,26,28–37]. For example, the model developed by Fischer and colleagues [30] also covers the problem identification, that is, the analysis of the underlying phenomenon before the generation of hypotheses. The skill “generating hypotheses” is divided into “questioning” and “hypothesis generation”. Moreover, “testing hypotheses” is subdivided into a construction phase (cf. the different notions of planning in Figure 1) and “evidence generation,” which is comparable to experimentation and observation [26,28,31]. The last facet of Klahr and Dunbar’s [29] model can be divided into “evidence evaluation,” such as the preparation and interpretation of experimental findings [26,32], and “drawing conclusions”. “Communicating and scrutinizing” completes the facets [30], whereas, for example, in Kambach’s model [26], communication is viewed as a skill relevant to all processes involved in SR. Some models also specify the documentation [28] skill as a separate facet, while this is seen as a cross-process skill relevant to all other facets in other models [26]. Recently, modelling skills have also been proposed to amend skills in conducting investigations to cover a broader notion of SR [10,33,34].

Klahr & Dunbar	generating hypotheses		testing hypotheses			analyzing evidence			
Mayer; Wellnitz et al.; Grube; Kunz	question	hypothesis	investigations			data analysis			
Hartmann et al; Krell et al.	formulating questions	generating hypotheses	planning investigations			interpreting data (= analyzing data and drawing conclusions)			
Nawrath et al.	formulating questions	generating hypotheses/assumptions	planning experiments	setting up experiments	observing/ measuring/ documenting	preparing data	drawing conclusions/ discussing		
Fischer et al.	problem identification	questioning	hypothesis generation	construction and redesign of artefact		evidence generation	evidence evaluation	drawing conclusions	communicating and scrutinizing
Kambach	phenomenon	question	hypothesis	planning	conducting experiments		interpreting		
Kambeyo	identifying variables	formulating questions	generating hypothesis	planning of variables	experimental plans		drawing conclusions		
Yanto et al.	problem formulating	problem exploration		investigation			conclusion		
Chang et al.	presenting question	presenting hypothesis	experimenting and data gathering			data analyzing, interpreting and concluding			

Figure 1. Overview of some models for scientific reasoning competency subdomain “conducting scientific investigations” (compiled from [10,24,26,28–37]).

For this study, part of Nawrath et al. and Kambach’s models [26,28] were combined to form a fine-grained model for measuring SR competency of preservice chemistry teachers. The skills *formulating research questions, generating hypotheses, planning experiments, observing and measuring, preparing data for analysis, and drawing conclusions* were included. In contrast to Kambach, the ability to specify the *phenomenon* is not assessed, but nevertheless, a phenomenon underlies the test instrument: A phenomenon is presented to the participants in an introduction and thereafter worked on. Furthermore, *setting up experiments* (cf. [28]) is not applied in our work, as this skill does not play a central role in scientific reasoning in chemistry and partly overlaps with practical work skills [24]. Documentation is understood

as an overarching skill following Kambach [26] and is thus not part of the skill *observing and measuring*, which is contrary to the model by Nawrath et al. [28].

1.2. Assessment of Scientific Reasoning

SR competency can be assessed with domain-specific as well as domain-general instruments in different test formats, such as paper-pencil tests or experimental tests. For the latter, real or virtual/mental experiment can be used [25]. Both types can employ multiple-choice items [10,34] or open formats [27] and assess theoretical notions of SR, performance-related measures [26,27], or self-assessments [35]. Some studies also use mixed formats [25]. The different methods for eliciting these skills have different properties regarding time consumption, practicability, individual diagnostics, external influences, congruence, and simultaneity [27]. Performance tests showed significantly larger effect sizes than multiple-choice tests [38]. Moreover, as Shavelson [9] pointed out, multiple-choice assessment can hardly be seen as a situation closely representing real life. Furthermore, multiple-choice questions may measure knowledge that the participants might be able to state but might not be able to put to practice (inert knowledge, see for instance [39]). Hence, a performance-based assessment format seems to be more closely tied to the competency to be inferred. However, performance-based assessment can be time-consuming for participants and researchers alike. In addition, there are differences in measuring skills in individual and group work: while group work enhances communication and therefore makes thoughts accessible to the researcher [40], better performance in problem solving of groups compared to individuals has been demonstrated [41], which may limit reliability of assessment of individual skills in group work situations.

Adults, and therefore also preservice teachers, experience difficulties in dealing with SR tasks: they tend to design confounded experiments or to misinterpret evidence to be able to verify their beliefs [1]. Kunz [36], Khan, and Krell [42] as well as Hartmann et al. [10] found higher SR competency in preservice science teachers with two natural science subjects, while this made no difference in a study conducted by Hilfert-Rüppell et al. [43]. Furthermore, students' SR competency differs by school type and progression in university studies [10,34,42]. Kambach's [26] findings suggest that preservice biology teachers either are very apt in describing phenomena, generating hypotheses, and interpreting results or do not show these processes at all. As for the other processes, skills show more variation among the sample. However, students also lack experimental precision and demonstrate deficient reasoning for their choice of material in planning investigations. While conducting experiments, they tend not to consider blanks and hardly ever plan intervals or end points while measuring. Finally, they tend not to prepare their data for analysis or refer back to their hypotheses while interpreting. Overall, Kambach's sample demonstrates variation of SR competency across the entire scale [26]. Hilfert-Rüppell et al. [43] demonstrated that preservice science teachers' SR skills *generating hypotheses* and *planning investigations* are deficient. However, they found that students' skill in *planning investigations* is moderated by their skill to *generate hypotheses*.

1.3. Learning Activities Supporting the Formation of Scientific Reasoning Competency

While the empirical and pedagogical literature has to offer various ideas and propositions for incorporating scientific inquiry into learning environments in schools and universities [44–46], preservice science teacher education still lacks inquiry learning activities [5,32]. For instance, lab courses mainly employ cook-book experiments [26,32,47]. If at all, the prospective teachers come into contact with forms of inquiry learning only in teaching methods courses in graduate education [10] or, if offered, while working or training in school laboratories [26]. Khan and Krell [42] therefore suggested a combination of contextualized, authentic scientific problem solving and its application to new contexts with tasks to reflect on problem solving and scientific reasoning on a meta level.

Still, the laboratory already "is the place of information overload" [48] (p. 266). Traditional cook-book experiments demand of the students to conduct, observe, note—and,

hopefully, also interpret and understand—an enormous number of elements [48]. However, most of these demands are clearly stated in the laboratory instructions. Working on a problem-solving experiment, students need to perform a similar amount of tasks as well as additional cognitive activities, such as understanding the problem and devising their own strategy to solving it. Since these demands occupy working memory space not directed to learning, especially open inquiry is seen to be ineffective due to cognitive overload [49]. Furthermore, unfamiliarity with the method of problem-solving experiments from previous laboratories might add to these strains. Reducing the amount of cognitive demands students have to face simultaneously can be achieved by scaffolding the problem-solving process [50], providing learners with worked examples [51,52], examples before the problem-solving process [53,54], or structuring tasks [55,56]. For instance, Yanto et al. [32] found that structuring three subsequent experimental classes using the three main types of inquiry (structured, guided, and open inquiry, cf. [6]) in a stepped sequence fosters preservice biology teachers' SR skills better than a traditional cook-book approach.

While the use of problem-solving and inquiry activities is widely seen as important, time-consuming learning activities like these do not always fit into tight schedules in schools and universities [44]. However, students may benefit from instruction before inquiry activities [57]. In a meta-analysis on the control of variables strategy, Schwichow et al. [38] showed that larger effects were achieved when learners were given a demonstration. Regarding chemistry laboratories, implementing instruction, such as demonstrations or examples as prelab learning activities, seems to be a promising approach [48]. This may be achieved by using educational videos [58–60].

1.4. Educational Videos as Pre Laboratory Activities

Educational videos are seen as a suitable medium to enhance students' preparation in undergraduate chemistry, for instance, regarding content learning in organic chemistry [61], calculations for laboratory courses [62], as well as the use of laboratory equipment and procedures [62–66]. Methods for the development of effective videos are subsumed in [67]. Cognitive load theory (CLT) [68] and cognitive theory of multimedia learning (CTML) [69] inform design of effective videos. For instance, exclusion of unnecessary details helps keep students' working memory from overloading (CLT), and making use of the visual and auditory channel in a way to avoid redundancy contributes to the effective use of both channels in educational videos (CTML). In terms of learning outcomes, Pulukuri et al. and Stieff et al. demonstrated that students preparing with videos statistically outperform a control group without any preparation [61] or an alternative treatment group preparing with a lecture [66]. However, many studies only report significant effects regarding the affective domain: students perceive videos to be helpful for preparation of and participation in the laboratory [60,62–64], even if no evidence can be found for their effectiveness on student performance [62,63]. Moreover, videos are still seen as rather new and motivating media in university education [70] and may therefore, like other newly advancing educational technologies, enjoy a novelty effect when used over a short period of time [61,70–73]. This may lead to an overestimation of their impact on student performance [61].

1.5. Self-Concept of Ability and Performance

“The relationship between self and performance is associated with an improvement in ability” [74] (p. 132). Self-concept is not a unidimensional construct but consists of various facets, such as academic self-concept [75]. Regarding students in an introductory chemistry course at university, House [76] showed that students' academic self-concept is a better predictor of first-year achievement in chemistry than, for example, grade of college admission test. Moreover, facets of self-concept can be broken down further, i.e., academic self-concept can be further differentiated for different subjects [75]. For example, Atzert and colleagues demonstrated that self-concept of ability can be measured regarding science experimentation [77]. Sudria and colleagues [78] compared self-assessment and objective assessments of preservice chemistry teacher students' practical skills in a chemistry labora-

tory. Their findings suggest that both students' self-assessed skills at the beginning and during the course correlate with objective assessment of their performance by the lecturer. Self-concept of ability usually is assessed with regard to three different norms: individual (i.e., development of abilities over time), social (i.e., own ability in relation to others), and criterial (i.e., own ability with regard to an objective measure) [79]. However, in agreement with the criterial rubric Sudria et al. used, Atzert et al. showed that only the criterial norm informs school students' self-concept of ability regarding science experimentation [77,78].

The aim of the project underlying this paper is both to foster preservice teachers' SR competency by implementing a small number of problem-solving experiments and explanatory videos into an already-existing lab course and to measure a potential increase in SR competency. This paper first describes an instrument for objectively measuring SR skills as well as a self-assessment questionnaire in which students rate their SR skills with regard to the criterial norm before and after the intervention. Using data from the pilot study, a first insight is given into development of students' SR skills.

2. Materials and Methods

2.1. Redesigning an Organic Chemistry Lab Course

Bearing in mind the insights from research on scientific reasoning and problem solving, we chose a 90-hour (3 credit points) organic chemistry lab course for second-year bachelor students [80] and redesigned 8 experiments into inquiry experimental problems cf. [6,46,81]. The intervention constituted approximately 30% of all lab course activities. To account for the high complexity of a full problem-solving process, each experiment was designed to focus mainly on one SR skill; *planning experiments* was further subdivided into (a) planning experiments (general aspects), (b) using the control of variables strategy, and (c) using blanks. Control of variables is central to the SR skill *planning experiments* [38]; however, students might not be familiar with this strategy (cf. Section 3.1). Using blanks is a specific form of controlling variables; yet, due to its application in analytical chemical, problems might be more familiar to second-year students than control of variables strategy. Moreover, using blanks (i.e., negative and positive controls) does not only cover the experimental design but addresses validity since it involves an examination of the method by (1) testing functionality of the reagents and (2) determining the limit of quantification [82–84]. Therefore, a distinction between using blanks and using the control of variables strategy was made.

For the lab course, this resulted in one experiment for each of the following skills: formulating research questions, generating hypotheses, planning experiments: general aspects, planning experiments: using the control of variables strategy, planning experiments: using blanks, observing and measuring, preparing data for analysis, and drawing conclusions. Students worked on the experiments in a stepped fashion: each consecutive experiment demanded of them to apply one more skill. Since formulating research questions and generating hypotheses are known to be more challenging to students than designing experiments and interpreting data [42,43], we organized the problem-solving experiments in a sequence from less to more challenging, starting with drawing conclusions in the first experiment up to the application of all skills in the final experiment [20]. Prior to the lab activity, each skill was explained and demonstrated to the students in a video lesson using examples different from those of the respective lab experiment. For instance, criteria for the generation of good scientific research questions or hypotheses were presented and applied to examples. In addition, students attended a colloquium on each experiment, discussing safety issues as well as specifics regarding experimental procedures and explanations with a lab assistant. In the redesigned course, the colloquium was also used to have students reflect on the content of each video lesson, i.e., students were asked to reproduce the main ideas taught in the video lesson and to apply them to the respective experiment. For example, they formulated their own research questions or presented their experimental planning. In the lab, students worked in pairs or groups of three if total participant count

was odd. They handed in lab reports after the course. Details on the redesigned lab are reported elsewhere [81,85].

Two cohorts served as control groups. They received the organic chemistry laboratory as originally designed, i.e., without an explicit focus on inquiry experiments. Students were neither asked to formulate research questions, generate hypotheses, plan their own experiments, nor draw conclusions with regard to a hypothesis. Instead, they were given cook-book descriptions of the processes to be conducted. If applicable to the experiment, students were only asked to choose from a given set of qualitative tests (such as Schiff test or Tollens reagent) and to conduct blanks for comparison of test results. They were not given any of the video lessons nor provided with any information from the video lessons in the colloquiums. To account for motivational effects of video media [61,70], the study also used an alternative treatment group. This group received the redesigned lab course with problem-solving experiments but watched videos about practical laboratory skills [62–66], i.e., their videos were unrelated to SR skills.

2.2. Hypotheses

The overarching goal of our project was to determine whether the redesigned lab course helps in fostering SR competency. Therefore, we adapted an already validated test instrument for school students [27] to use with preservice chemistry teachers and complemented it with a self-assessment questionnaire. Psychometric properties were examined in a pilot study, and the following hypotheses were tested to account for the purposefulness of Kraeva's instrument [27] for our target group:

Hypothesis 1. *In the adapted version of the test instrument, accompanying variables (prior knowledge, methodological knowledge, documentation skill) correlate in the same pattern as in Kraeva's [27] findings.*

Hypothesis 2. *Students in the control group score similar points in accompanying variables (prior knowledge, methodological knowledge, documentation skill) in pre- and post test since both test booklets are expected to be comparable, as they do not require prior knowledge [27].*

Since traditional cook-book labs should already support some SR skills also associated with cook-book experiments, such as *observing and measuring* or *preparing data for analysis* [6,86,87], a control group was used to determine the extent to which the cook-book lab already fosters SR skills. Since both treatment groups worked on the problem-solving experiments in the lab, these were both expected to gain SR competency over the course of the lab. Nevertheless, the treatment group watching the SR-related videos (SR group) received more support in structuring the problem-solving process than the alternative treatment video group that watched SR-unrelated videos (alternative group). Therefore, the SR group was expected to benefit more from the lab course, which should manifest itself in a greater learning gain [61,66]. We hypothesized as follows:

Hypothesis 3. *Students in the control group show an increase in SR skills observing and measuring from pretest to posttest (i.e., after participation in the traditional lab course) but not in skills generating hypotheses, planning experiments, or drawing conclusions.*

Hypothesis 4. *Students in both treatment groups (SR group and alternative group) show an increase in SR competency from pretest to posttest.*

Hypothesis 5. *Students in the SR group show a greater learning gain in SR competency than students in the alternative group.*

2.3. Data Collection

Following Shavelson's [9] requirements for competency measurement, SR competency can be inferred from measuring a set of complex skills (such as formulating hypotheses, planning experiments, drawing conclusions, see Section 1.1) observable in a performance situation (experimental problem-solving tasks in the test instrument) close to a real-world situation (such as the problem-solving experiments in the laboratory). Tasks and scoring manuals need to be standardized for all participants (as presented below) and yield a score for the level of performance from which competency can be inferred. Moreover, the skills measured are supposed to be improvable through teaching and practice (that is, by the students attending a laboratory course such as the intervention presented) as well as dependent on disposition (such as self-regulation due to self-assessment in the respective skills). Therefore, we chose to build on an already validated, qualitative instrument with which the procedural structures of students' problem-solving processes in an inquiry experiment can be determined using video recordings and written records [27]. Processes observed by Kraeva [27] were *generating hypotheses*, *planning experiments*, and *drawing conclusions*. Since Kraeva's instrument was validated with high school students grade from 5 to 10, we report here the adaptation to the target group of university students. Using an expanded manual, the following SR skills were measured in the pilot study: *generating hypotheses*, *planning experiments*, *observing and measuring*, and *drawing conclusions*. Additional tasks assessing the skills *developing questions* and *preparing data for analysis* were constructed for the main study. Due to the pandemic, data on the latter two tasks so far could only be collected on five participants. Therefore, only data on the first four skills are presented here. As accompanying variables, prior knowledge, documentation skills, and methodological knowledge were assessed using Kraeva's [27] instrument and manual. After the performance test, the students filled out a self-assessment questionnaire in which they assessed their own SR skills (*developing questions*, *generating hypotheses*, *planning experiments*, *using control-of-variables strategy*, *using blanks*, *observing and measuring*, *preparing data for analysis*, and *drawing conclusions*) on a five-point-scale. In addition, demographic data, such as age, gender, and parameters for students' learning opportunities in chemistry (subject combination, semesters spent at university, and success in organic chemistry), were collected.

The test was administered in German with standardized test instructions before (pretest) and after completion of the lab course (posttest) with two similar test booklets on different chemical topics (adapted from [27]). The survey usually took place in pairs so that conversations could be recorded while videotaping. Data were collected anonymously with cameras positioned to only film participants' hands and working surfaces. Students who did not wish to be videotaped were seated at a table without recording equipment. All students who participated in the study signed consent forms. Research procedures were in accordance with ethical standards of Technische Universität Braunschweig. Participants of cohort 2020 were recorded individually because of pandemic regulations. Nevertheless, due to using a think-aloud protocol (adapted from [88]), it was still possible to capture students' thoughts.

2.4. Description of the Test Booklet

Paper-pencil tests started with two tasks on prior knowledge (Task 1: everyday knowledge and Task 2: chemical knowledge [27] (p. 81)) regarding the respective topic (e.g., surface tension). Students were then shown a slow-motion video of the phenomenon to be investigated and asked to document experimental procedures and observations (Task 3). This was followed by a videotaped sequence of tasks ("Experimental Tasks", see Figure A1a and [27] (p. 71)), including a problem-solving experiment: first, students were asked to generate a hypothesis (hypothesis I) about the phenomenon in the slow-motion video (Experimental Task a). They were then asked to plan and conduct an experiment related to their hypothesis (experiment I), document the procedures, and draw a conclusion (Experimental Task b). This sequence was videotaped to observe students'

actual problem-solving process since protocols are known to not necessarily contain all steps discussed or conducted but rather a selection of those procedures that students judge worth reporting [89]. After the experimental task, students were asked to give an explanation for the phenomenon observed earlier in the slow-motion video using findings from their own experiments (Task 4; [27] (p. 82)). In Task 5 (see Figure A1b), they were asked to think of other conditions that affect the phenomenon and to develop a new research question and a corresponding hypothesis (hypothesis II). They then planned another experiment to test their hypothesis (experiment II), this time, however, without conducting it. Finally, a method for measuring surface tension was described to the students; they were given measurement data and asked to prepare a diagram for future analysis (Task 6, see Figure A2) as well as to extract information from the diagram to answer a question on data analysis.

2.5. Coding Manuals for Scientific Reasoning Skills

Transcriptions of the videotaped sequence and students' written answers in the test booklet were analyzed using a coding manual. Accompanying variables were assessed in individual work from students' written records. Table 1 shows which data sources were taken into account for analyses of SR skills.

Table 1. Data sources for analyses of SR skills and accompanying variables.

SR Skill	Written Records	Video Transcripts
developing questions	individual work	-
generating hypotheses I	individual work	discussion in pair work
generating hypotheses II	individual work	-
planning experiments I	-	pair work
planning experiments II	individual work	-
observing and measuring	-	pair work
preparing data for analysis	individual work	-
drawing conclusions	individual work ¹	-

¹ Individual documentation of findings from experimental tasks in pair work.

Coding manuals for rating students' SR skills were deductively developed from the literature [26–28,90] and inductively complemented with data from the control group. For calculation of interrater reliability, 13% of video transcripts and written records were coded by two raters (author 1 and a trained student research assistant, see Section 3.3). Students' skills were assessed on four-point scales using the manual (see Table A1), whereby a full score indicated that the skill is fully developed. For example, there are four expressions for the skills *observing and measuring* as well as *planning experiments I/II* (e.g., no experiment, explicating planning, plan not suitable, plan suitable). The rating of the other skills is divided into four categories (e.g., for *generating hypotheses I/II*: no hypothesis, hypothesis, explanation, relationship). Points can be awarded independently of each other. The highest expression shown by a student was coded even if the same student did not demonstrate the respective level at another occasion in the task because it was assumed that once competency is expressed in performance, it can, in principle, be shown again and again.

2.6. Sample

Sixty preservice chemistry teacher students participated in the pilot study. Students were on average 22.1 years old (SD = 3.2). The majority identified as female (34 participants), 20 participants as male, and 4 did not provide a gender identification. Ratio of female to male students is usually high at Technische Universität Faculty of Humanities and Education Studies [91]. Average grade of school leaving certificate was 2.3 (SD = 0.55; "Abitur", grades may vary from 1.0 to 4.0, with 1.0 being the best possible grade). Forty participants studied two STEM subjects, and 14 participants studied chemistry in combination with a non-STEM subject. The majority of the preservice teacher students in the sample

planned to teach at secondary schools up to 12th grade ($n = 42$; German “Gymnasium”), and 15 participants planned to teach at secondary schools up to 9th or 10th grade (German “Realschule/Hauptschule”). On average, participants were in their 3rd semester of the bachelor (IQR = 2.75) when attending the organic chemistry laboratory. In addition, most participants ($n = 50$) had attended the corresponding lecture in organic chemistry before the laboratory; 29 had also passed the respective exam.

The data were collected between 2017 and 2020 in a pretest-posttest design, i.e., immediately before and after the laboratory course. Across all cohorts, some students refused videography, and thus, in some cases, less data are available for skills *planning experiments I* and *observing and measuring* than for those skills assessed from written records (see Table 2). The self-assessment questionnaire was not administered in the control group because it was not added to the study design until production of explanatory videos was completed. SR tasks assessing *formulating research questions* and *preparing data for analysis* were piloted in 2020 with a small cohort due to the pandemic.

Table 2. Sample sizes for tasks A1 to A5 (accompanying variables, [27]) and SR skills by group in the pretest.

Group	A1 to A5	H I	H II	P I	P II	OM	C	Q	D	SA
control	28	28	28	11	28	11	28	– ¹	– ¹	– ¹
SR	18	18	18	17	18	17	18	6	6	18
alternative	14	14	14	11	14	11	14	– ¹	– ¹	14
Sum total	60	60	60	39	60	39	60	6	6	32

Abbreviations: H, generating hypotheses; P, planning experiments; OM, observing and measuring; C, drawing conclusions; Q, formulating research questions; D, preparing data for analysis; SA, self-assessment. ¹ Task/questionnaire not yet implemented.

3. Results

Psychometric properties of the instruments were calculated using pretest data from the pilot study. Hypotheses were tested using pre- and posttest data. For a first insight into the effects of the newly designed laboratory, pre- and posttest measurements from the treatment groups were examined.

3.1. Self-Assessment Questionnaire

Self-assessment data on SR skills were collected in the treatment groups, totaling 32 participants. Item parameters of the self-assessment instrument show that students rated their initial abilities as rather high (M from 3.06 to 4.50, see Table 3; rating on a scale from 1 to 5). However, the majority of participants already judged their skills in *using blanks*, *observing and measuring*, and *drawing conclusions* to be very high before participating in the laboratory. Interestingly, for item “*using control of variables strategy*,” 15 out of 32 students chose the alternative answer “I don’t know,” resulting in only 17 valid answers. Item *planning experiments* was answered by six participants because it had only been added to the questionnaire in 2020.

Both items *planning experiments* and *using control of variables strategy* were excluded from calculations due to the small number of answers. Exploratory factor analysis of the remaining six items using principal component analysis (PCA) with varimax rotation showed a two-factor solution judging by Kaiser criterion (see Table 4) [93]. Rotated component matrix of the two-factorial structure indicated that the second factor consisted of only two items, *using blanks* and *observing and measuring*. These items identified as measuring skills not exclusively associated with inquiry experiments but also needed when conducting cook-book experimental procedures. Both items were excluded from the potential scale of SR skills. Reliability of the adapted four-item scale “self-assessment of scientific reasoning competency” gave an acceptable Cronbach’s α of 0.787 (4 items, $n = 29$) and a rather high inter-item correlation of 0.488 but still considerably lower than α [94,95]. No improvements of Cronbach’s α were achievable by further removal of items. Mean score on the four-item scale was $M = 14.97 \pm 2.442$.

Table 3. Self-assessed scientific reasoning skills in the pretest (M, mean; SD, standard deviation; P_i , item difficulty; n , sample size; assessment on a 5-point scale).

SR Skill	M	SD	P_i	n
formulating research question	3.27	0.907	0.57	30
generating hypotheses	3.68	0.832	0.67	31
planning experiments ¹	3.83	0.753	0.71	6
using control of variables strategy	3.06	1.298	0.51	17
using blanks	4.09	0.818	0.77	32
observing and measuring	4.50	0.672	0.88 ²	32
preparing data for analysis	3.88	0.833	0.72	32
drawing conclusions	4.03	0.647	0.76	32

¹ Item removed from the scale due to small sample size. ² Item difficulty indicates ceiling effect. (A ceiling effect is defined as “a situation in which the majority of values obtained for a variable approach the upper limit of the scale used in its measurement. For example, a test whose items are too easy for those taking it would show a ceiling effect because most people would achieve or be close to the highest possible score. In other words, the test scores would exhibit skewness and have little variance, thus prohibiting meaningful analysis of the results” [92].)

Table 4. Rotated component matrix of self-assessed scientific reasoning skills in the pretest ($n = 29$). Factor loadings negligibly small (<0.3) [93] are set in gray.

SR Skill	Component 1	Component 2
formulating research question	0.798	−0.053
generating hypotheses	0.755	0.247
using blanks	0.074	0.879
observing and measuring	0.115	0.900
preparing data for analysis	0.809	−0.010
drawing conclusions	0.745	0.261

Pre- and posttest data of self-assessment from the treatment groups were compared using the four-item-scale. Wilcoxon test was used due to small sample sizes. In both groups, a tendency for improvement toward the posttest is visible yet not significant (see Table 5).

Table 5. Analysis of differences between pretest and posttest in SR skills in the treatment groups, calculated using Wilcoxon test (M, mean; SD, standard deviation; n , sample size; Z , parameter of Z-distribution; p , significance level).

Treatment Groups	Pre			Post			Z	p ¹
	M	SD	n	M	SD	n		
alternative	15.75	2.527	12	16.75	2.137	12	−1.299	0.116
SR	14.44	2.366	16	15.00	2.608	16	−0.829	0.215

¹ Exact significance is reported due to small sample size ($n < 30$).

3.2. Accompanying Variables

As Kraeva [14] had constructed the instrument for school students grade 5 to 10, yet as a tool not relying on prior knowledge, we investigated whether item difficulties in the accompanying variables (tasks A1 to A5) might hint at ceiling effects [92], potentially rendering the test too easy for university students. Means and item difficulties from pretest data of the pilot study show that students achieve moderate to high scores in the accompanying variables of the paper-pencil test. The test does not produce ceiling effects except for the task on content knowledge (see Table 6).

Kraeva [27] found small but significant correlations between tasks A1 (prior knowledge from everyday life) and A2 (prior knowledge from chemistry content knowledge) as well as between tasks A4 (explaining results) and A5 (generating a hypothesis and planning a corresponding but hypothetical experiment) and had hence subsumed tasks A1 and A2 to form a measure for prior knowledge and tasks A4 and A5 to measure methodological

knowledge. We therefore expected to find similar correlations, while task A3 (“documentation”) was expected not to correlate (Hypothesis 1). Correlations with medium effect sizes [96] were found between tasks A1 and A2 ($r = 0.332$; $p = 0.010$; $n = 60$), tasks A3 and A1 ($r = 0.320$; $p = 0.013$; $n = 60$), and tasks A3 and A4 ($r = 0.301$; $p = 0.019$; $n = 60$), but no significant correlations with task A5 were found. For following analyses, tasks A1 and A2 were therefore subsumed as “prior knowledge” [27]; tasks A3, A4, and A5 were treated as separate items. In addition, task A5 was also rated using coding manuals for SR skills *generating hypotheses II* and *planning experiments II* (see Section 3.3).

Table 6. Item parameters of tasks A1 to A5, pretest data from the pilot study (M, mean; SD, standard deviation; P_i , item difficulty; n , sample size).

Task	M	SD	P_i	n
Everyday life knowledge (A1)	1.32 ¹	0.833	0.66	60
Content knowledge (A2)	1.62 ¹	0.691	0.81 ³	60
Prior knowledge (A1 + A2)	2.93 ²	1.247	0.73	60
Documentation skill (A3)	0.88 ¹	0.904	0.44	60
Explaining (A4)	1.25 ¹	0.795	0.63	60
Hypothesis and planning (A5)	0.90 ¹	0.730	0.45	60
Methodological knowledge (A4 + A5)	2.15 ²	1.147	0.54	60

¹ Maximum of 2 points. ² Maximum of 4 points. ³ Item difficulty indicates ceiling effect [92].

Furthermore, we expected that pre- and posttest performance of participants would not differ in the accompanying variables, accounting for comparability of the pre- and posttest booklets (Hypothesis 2). To eliminate any potential influence from the intervention, only data from the control group were used in the comparison. Table 7 shows results from Wilcoxon signed-rank test indicating no significant differences between pretest and posttest performance of the control group in the accompanying variables.

Table 7. Analysis of differences between pretest and posttest in tasks A1 to A5 in the control group, calculated using Wilcoxon signed-rank test (M, mean; SD, standard deviation; n , sample size; Z , parameter of Z-distribution; p , significance level).

Task	Pre			Post			Z	p ³
	M	SD	n	M	SD	n		
Prior knowledge (A1 + A2)	3.18 ¹	1.278	28	3.07 ¹	1.016	28	−0.149	0.893
Documentation skill (A3)	1.00 ²	0.861	28	0.89 ²	0.875	28	−0.528	0.637
Explaining (A4)	1.29 ²	0.763	28	0.93 ²	0.813	28	−1.586	0.132
Hypothesis and planning (A5)	0.93 ²	0.766	28	1.04 ²	0.693	28	−0.786	0.515

¹ Maximum of 4 points. ² Maximum of 2 points. ³ Exact significance is reported due to small sample size ($n < 30$).

3.3. Scientific Reasoning Skills

In the pilot study, students’ SR skills *generating hypotheses I/II*, *planning experiments I/II*, *observing and measuring*, and *drawing conclusions* were assessed on four-point scales using either their written records of tasks in individual work, such as *generating hypotheses II* or *drawing conclusions*, or video transcripts of tasks in pair work, such as *planning experiments I* or *observing and measuring* (see Table 1 for details on data source per skill). Since analysis of accompanying variables (see Section 3.2) showed that tasks A4 and A5 did not correlate as was found by Kraeva [27], Task A5 was used to assess students’ SR skills *generating hypotheses II* and *planning experiments II* in individual work with the manuals presented in Section 2.5 (Table A1).

Content validity of the instrument and manuals was established in a group discussion of eight members of staff in chemistry- and biology-teaching methodology. Reliability of the data collection was assessed by computing interrater reliabilities for the manuals. For this, the author conducted a rater training with the second rater (student research assistant) after the development of the manual, in which the manual was first presented

in general and discussed using some examples. Finally, the student raters' questions were clarified. Afterwards, the second rater coded 13% of the material and noted further questions and ambiguities, which were then clarified in a second rater training session. This was followed by the final coding of the material by both raters (13% student rater, entire data set author 1), from which the results of the ICC were computed. Intraclass correlation for absolute rater agreement in the presence of bias (ICC (A,1); [97]) was calculated, yielding excellent reliabilities ranging from 0.915 to 1.000 ([98]; see Table 8). Interrater reliabilities for the newly developed tasks measuring SR skills *formulating research questions and preparing data for analysis* were not yet calculated because of small sample sizes in 2020 due to pandemic regulations.

Table 8. Interrater reliability (ICC (A,1); [97]) of the manuals for SR skills.

SR Skills	ρ	p
generating hypotheses	0.915	0.000
planning experiments	1.000	0.000
observing and measuring	0.968	0.000
drawing conclusions	0.971	0.000

Means of students' scores in the SR skills indicate that students already achieved moderate results in the pretest (see Table 9). Item difficulties were high but showed no ceiling effects, indicating that the tasks were not too easy for university students. For those skills that were both assessed in individual work (*generating hypotheses II, planning experiments II*) and in pair work (*generating hypotheses I, planning experiments I*), item parameters indicated that pair work assessment results in a higher item-difficulty value, i.e., tasks in pair work are easier for the students than tasks in individual work. Exploratory factor analysis (PCA, varimax rotation [93]) indicated a two-factorial structure judged by Kaiser criterion (see Table 10). Rotated component matrix showed that component 1 represents skills assessed in individual work using written records (*generating hypotheses II, planning experiments II and drawing conclusions*), and component 2 represents skills assessed in pair work using video data (*planning experiments I and observing and measuring*) as well as written records (*generating hypotheses I*). However, *observing and measuring* shows a negative loading and was therefore excluded.

Reliability was calculated for the potential scales "individual SR competency" using variables *generating hypotheses II, planning experiments II and drawing conclusions* as well as "SR competency in pair work" using variables *generating hypotheses I and planning experiments I*. Cronbach's $\alpha = 0.578$ was found to be rather low for the three-item scale "individual SR competency" ($n = 60$) but considerably higher than moderate average inter-item-correlation of 0.335 [95]. For the two-item-scale in pair work, Cronbach's $\alpha = 0.292$ was not acceptable ($n = 39$). Even though some authors argue that a Cronbach's α lower than 0.7 is acceptable if item content is meaningful [94,95], we decided not to use the scales but to report analyses of SR skills item-wise.

Table 9. Item parameters of objectively assessed scientific reasoning skills, pretest data from the pilot study (M, mean; SD, standard deviation; P_i , item difficulty; n , sample size). From 0 to 3 points were achievable in each skill.

SR Skills	M	SD	P_i	n
generating hypotheses I	2.15	0.880	0.72	60
generating hypotheses II	1.67	0.816	0.56	60
planning experiments I	2.13	0.656	0.71	39 ¹
planning experiments II	1.67	1.100	0.56	60
observing and measuring	2.03	0.668	0.68	39 ¹
drawing conclusions	2.07	1.023	0.69	60

¹ Sample size is smaller because not all participants agreed to the videography.

Table 10. Rotated component matrix of objectively assessed scientific reasoning skills in the pretest ($n = 39$). Factor loadings negligibly small (<0.3) [93] are set in gray.

SR Skills	Component 1	Component 2
generating hypotheses II	0.852	0.039
planning experiments II	0.777	−0.146
drawing conclusions	0.770	0.120
generating hypotheses I	0.213	0.784
planning experiments I	−0.288	0.508
observing and measuring	−0.023	−0.685

Table 11 shows mean pretest and posttest scores for SR skills of all three groups. On average, students in all groups showed moderate abilities in all skills as well as a tendency for increase in the posttest in nearly all skills. Participants from the alternative treatment group seemed to achieve higher performances in pair work in the pretest (cf. *generating hypotheses I, planning experiments I*). To determine whether the traditional laboratory already enhances students' SR skills (Hypothesis 3), two cohorts served as control groups. They received the organic chemistry laboratory as originally designed, i.e., without an explicit focus on inquiry experiments. Pre- and posttest data from the control group were tested for differences in the variables *generating hypotheses I/II, planning experiments I/II, observing and measuring, and drawing conclusions*. Differences between pre- and posttest were only found to be statistically significant for the skill *observing and measuring* (see Table 11). In this skill, the posttest shows a ceiling effect, as all participants achieved the full score in the posttest. Hence, the posttest may have been too easy for the participants of the control group regarding this skill.

Table 11. Comparison of pretest and posttest mean scores in objectively assessed scientific reasoning skills, calculated using Wilcoxon signed-rank test (M, mean; SD, standard deviation; n , sample size; Z , parameter of Z -distribution; p , significance level). p -values for nonsignificant test results ($p > 0.05$) are set in gray.

SR Skills	Pre			Post			Z	p^1
	M	SD	n	M	SD	n		
			control group					
generating hypotheses I	2.04	1.055	27	2.41	0.694	27	−1.487	0.081
generating hypotheses II	1.63	0.792	27	1.59	0.931	27	−0.080	0.491
planning experiments I	2.36	0.809	11	2.82	0.405	11	−1.406	0.125
planning experiments II	1.59	1.047	27	1.81	1.145	27	−0.851	0.221
observing/measuring	2.00	1.095	11	3.00	0.000	11	−2.460	0.008
drawing conclusions	2.04	1.224	27	2.07	0.997	27	−0.054	0.485
			alternative group					
generating hypotheses I	2.46	0.776	13	2.46	0.877	13	−0.122	0.500
generating hypotheses II	1.46	0.877	13	2.15	0.689	13	−1.852	0.043
planning experiments I	2.45	0.522	11	2.64	0.809	11	−0.816	0.344
planning experiments II	1.62	1.121	13	2.77	0.832	13	−2.461	0.008
observing/measuring	2.09	0.701	11	1.91	0.701	11	−0.694	0.242
drawing conclusions	2.15	0.987	13	2.31	0.751	13	−0.491	0.375
			SR group					
generating hypotheses I	2.00	0.612	17	2.59	0.618	17	−2.352	0.014
generating hypotheses II	1.94	0.827	17	2.18	0.728	17	−1.069	0.216
planning experiments I	1.76	0.437	17	2.76	0.664	17	−3.127	0.001
planning experiments II	1.82	1.185	17	2.24	0.970	17	−1.137	0.146
observing/measuring	2.00	0.000	16	2.13	0.619	16	−0.816	0.344
drawing conclusions	2.06	0.748	17	1.71	0.985	17	−1.604	0.091

¹ Exact significances are reported due to small sample sizes ($n < 30$).

Furthermore, we hypothesized that both treatment groups would show an increase in SR competency (Hypothesis 4). Both groups had significantly higher mean scores in *generating hypotheses* and *planning experiments* in the posttest than in the pretest (see Table 11). Interestingly, for the SR group, this only applies to the skills assessed in pair work, while for the alternative group, the increase is only significant for the skills assessed in individual work. Regarding the alternative group, skills assessed in pair work were already rather high in the pretest compared to individual skills. If pretest values are high, there is less room for improvement. Nevertheless, a total number of 13 participants equal 6 groups at most, reducing validity of the comparison. Furthermore, in contrast to the control group, neither treatment group achieved a significant increase in *observing and measuring* toward the posttest. Neither the treatment groups nor the control group showed an increase in the skill *drawing conclusions*. Still, it should be noted here that the small sample sizes of the pilot study, especially in the alternative group, limit generalizability of these findings.

So far, performances of control and treatment groups were compared independently of each other, yielding five significant achievement gains. Hypothesis 5 assumed that students in the SR group show a greater learning gain in SR competency than students in the alternative group. To enable a comparison among the groups, gains in each skill were calculated by distracting participants' pretest scores from their posttest scores. Then, Kruskal–Wallis H test was performed on the pre-post differences, indicating the only group difference in the skill *planning experiments I* (see Table 12). A post-hoc test (with Bonferroni correction) showed that the group difference resulted only from a significant difference of learning gain between alternative group and SR group ($z = -2.487$; $p = 0.039$). So far, it can be concluded that control group and alternative group did not differ in learning gains, but participating in the SR group led to a significantly larger learning gain in the skill *planning experiments I*. Beyond that skill, no other differences were found between alternative and SR groups or control group and SR group. As was stated before, limitations regarding generalizability of these findings apply due to the small sample sizes.

Table 12. Comparison of groups for mean pretest-posttest differences in objectively assessed scientific reasoning skills, calculated using Kruskal–Wallis H test (M, mean of pre-post difference; SD, standard deviation; n , sample size; H, parameter of H-distribution; p , significance level). p -values for nonsignificant test results ($p > 0.05$) are set in gray.

SR Skills	Control Group			Alternative Group			SR Group			H (2)	p
	M	SD	n	M	SD	n	M	SD	n		
generating hypotheses I	0.37	1.214	27	0.00	1.414	13	0.59	0.870	17	1.455	0.483
generating hypotheses II	−0.04	1.224	27	0.69	1.251	13	0.24	0.903	17	2.393	0.302
planning experiments I	0.45	1.036	11	0.18	0.751	11	1.00	0.866	17	6.742	0.034 ¹
planning experiments II	0.22	1.340	27	1.15	1.281	13	0.41	1.326	17	3.675	0.159
observing/measuring	1.00	1.095	11	−0.18	1.250	11	0.13	0.619	16	5.705	0.058
drawing conclusions	0.04	1.255	27	0.15	1.214	13	−0.35	0.862	17	1.495	0.474

¹ Post-hoc test results: (z (control vs. alternative) = 0.746; $p = 1.000$; z (alternative vs. SR) = -2.487 ; $p = 0.039$; z (control vs. SR) = -1.664 ; $p = 0.288$).

4. Discussion and Limitations of the Study

In this study, an already validated, performance-based instrument for description of SR processes of school students was adapted for measurement of SR competency of preservice chemistry teachers. Accompanying variables adopted from Kraeva [27] as well as tasks measuring SR skills were found to be suitable for preservice chemistry teachers regarding difficulty and comparability of test booklets (Hypothesis 2). Kraeva's performance test originally only involved *generating hypotheses*, *planning experiments*, and *drawing conclusions* in a mixed format of individual and pair work. Due to the fact that, in contrast to Kraeva [27], no significant correlation between accompanying variables A4 and A5 could be identified, these variables were not summarized to form a measure for methodological knowledge but treated as separate items. Even though Hypothesis 1 therefore had to be rejected in parts, the data from A5 could now be used to assess SR skills

generating hypotheses II and *planning experiments II* from individual work. Furthermore, the test was extended to measure *observing and measuring* in pair work in the pilot study. Factor analysis indicated a two-factorial structure of SR skills, separating skills assessed in individual work from those assessed in pair work. This is in accordance with findings from other studies comparing individual and group performance [40,41]. Even though reliabilities were low, tasks assessing skills individually yielded slightly more reliable data. Still, excellent interrater reliabilities were found indicating reliability of the method for collecting data on SR skills. Hence, for use in the main study, new tasks assessing skills *formulating research questions* and *preparing data for analysis* were added to the test. Factor analysis of SR self-assessment items indicated that the skills *using blanks* and *observing and measuring* load on a different factor than the other SR skills, such as *formulating research questions* or *generating hypotheses*. The former two skills seem to be not only relevant to inquiry experiments exclusively but also to cook-book experimentation. For example, Sudria and colleagues included *observing* in a set of practical laboratory skills [78].

The second aim of this project was to enhance preservice chemistry teachers' SR competency through experimental problem solving and explanatory videos in an organic chemistry lab course. First insights can be inferred from comparison of control and treatment groups in the pilot study. Even though 60 students in total participated in the pilot study, the rather small sample sizes in each group still limit generalizability of the findings. Both SR self-assessment and objective assessment data show that preservice chemistry teachers in their second year at university already demonstrate substantial skill before attending the laboratory. That is, without having received any explicit instruction on inquiry learning or scientific reasoning so far. In comparison to instruments used with secondary preservice science teachers in other studies [10,34,42], the instrument presented here seems to be less difficult. This is in accordance with the origin of the instrument, which was originally developed for school students [27]. Nevertheless, increases in several skills were measurable (see below).

Students rated their own abilities in *using blanks*, *drawing conclusions* and *observing and measuring* as particularly highly developed, while lower self-assessment of skills was found for *formulating research questions* and *using control of variables strategy*. Especially the *control of variables strategy* may be unknown to some preservice teacher students in their second year of the bachelor, which might explain why students more frequently chose the alternative answer "I don't know" with this item. A similar pattern was found in the SR skills assessed from students' performance: students' individual performance was found to be relatively high in *drawing conclusions* and moderate in *generating hypotheses* and *planning experiments*. This is accordance with findings from Krell and colleagues as well as Khan and Krell that students' performances are lower in *formulating research questions* and *generating hypotheses* than in *planning investigations*, *analyzing data*, and *drawing conclusions* [34,42]. As was also demonstrated before [41], students scored more points in skills assessed in pair work than in individual work: Performance for *generating hypothesis I* and *planning experiments I* (assessed in pair work) tended to be higher than for *generating hypotheses II* and *planning experiments II* (assessed in individual work). This is in accordance with findings that groups have higher success in problem solving than individuals because they engage more actively in explanatory activities [41,99]. However, it cannot be said with certainty that the higher score is exclusively due to the work in pairs. Stiller and colleagues identified several features rendering test items difficult [100], such as text length, use of specialist terms, and cognitive demands, i.e., use of abstract concepts (for instance, also [101] in this special issue). A comparison of the experimental task and task 5 (see Figure A1) indicates no difference in text length or use of specialist terms. Tasks involving abstract concepts require participants to build "hypothetical assumptions [. . .] not open to direct investigation" [100] (p. 725). This only holds true for *planning experiments II* (planning of a hypothetical experiment) but not for *generating hypotheses I/II*, as these are both hypothetical tasks. Moreover, students were not observed to change their answers in task *generating hypotheses I* after writing up their answers to the experimental task. Kraeva's

test construction followed the Model of Hierarchical Complexity in chemistry (MHC) [102], which describes task complexity with regard to the number of elements to be processed and their level of interrelatedness. Both the experimental task and task 5 were constructed on the highest level (“multivariate interdependencies”) [102] (p. 168); therefore, both tasks require the same cognitive demands. Additionally, the students were videotaped while solving the experimental task, which may have led to greater care and effort in solving the task. Therefore, it cannot be conclusively clarified whether this is an effect of pair work.

Regarding the increase in skill achieved through participation in the lab course, some learning gains were found in the control and treatment groups. As was expected for the control group, an increase was only found for *observing and measuring* but not for *generating hypotheses, planning experiments, or drawing conclusions* (Hypothesis 3). This may be attributed to the fact that in the traditional laboratory, students were not asked to generate their own research questions or hypotheses. Hence, there was also no need for them to reason with respect to question or hypothesis, consequently yielding no increase in these skills [6,86,87]. Hypothesis 3 was therefore provisionally accepted. Increases in SR skills in the treatment groups were not as clear cut as hypothesized. Both treatment groups showed an increase in *generating hypotheses* and *planning experiments*, whereas no increase was found for *observing and measuring* and *drawing conclusions*. Thus, Hypothesis 4 could be provisionally accepted for the respective skills. As the control group’s skill in *observing and measuring* increased, an increase would have been expected in the treatment groups as well. This may be attributed to several possible reasons: On the one hand, cognitive demands placed on the treatment groups due to the additional and new learning objectives in the intervention (such as generating hypotheses) could have been too high, therefore reducing cognitive capacity directed at skills students might have perceived as already familiar to them. On the other hand, since *observing and measuring* showed negative factor loading on SR skills in pair work, there might as well be an issue with the assessment of this skill either in the manual or in the task. Hence, this skill should undergo revision before start of the main study. So far, Hypothesis 5 had to be rejected since the SR group only showed a significantly larger learning gain than the alternative treatment group in one skill, *planning experiments I*, but no difference in learning gain compared to the control group. Since the data analyzed here belonged to the pilot study and therefore only give a first indication of the effectiveness of the intervention, both hypotheses 4 and 5 will have to be tested again in the main study.

Qualitative assessment was chosen to arrange for a more individualized view on students’ skills; yet, quantitative analyses show that small sample sizes are a serious limitation of the presented investigation. This applied particularly to the very small sizes of the treatment groups due to the piloting. Resulting from this, issues arose for the ratio of items to participants in the factor analyses as well as regarding the low reliabilities of the instruments. Furthermore, conducting parts of the performance tasks in pair work led to reliability issues in comparison to individually assessed SR skills, negating the advantage of the pair-work format in enhancing communication and hence accessibility of participants’ thoughts [40]. In addition, since the original test instrument was constructed for school students, we expected preservice teachers to achieve moderate to high scores. In some variables however, this produced ceiling effects [92], such as in the control group’s posttest for performance-based SR skill *observing and measuring* or in the self-assessment in the respective skill. This may lead to a failure of the instruments in differentiating potential gains in these skills due to the treatment. Furthermore, negative factor loading of SR skill *observing and measuring* demands that the respective task and manual should undergo revision before conducting the main study. Regarding the late introduction of the self-assessment questionnaire in the design of the study, comparison of self-assessed skills in the control group was impossible. Moreover, no gains in self-assessment of skills were found for the treatment groups. It cannot be ruled out that the pandemic had an influence on student motivation in the 2020 cohort, as lab activities had to be conducted under strict pandemic restrictions, for example, prohibiting pair work in the lab. Furthermore, the

pandemic may also have had an impact on the researchers' and assistants' performance in the laboratory due to uncertainties in the planning process. Since the 2020 cohort was part of the SR group, this limits the scope of the findings from comparison of treatment groups even more. New pandemic regulations may also hinder the further conduction of this study.

5. Conclusions and Future Directions

With the performance-based instrument presented here, so far, four SR skills as well as gains in SR skills could be measured on a fine-grained level. Hence, the main aim of this pilot study was partially achieved. For the further course of the project, assessment tasks for the skills *developing research questions* and *preparing data for analysis* will undergo further investigation as soon as pandemic restrictions permit standardized test administration and delivering of the laboratory. In addition, task and manual assessing *observing and measuring* will be inspected critically. As for the self-assessment questionnaire, items for *planning experiments* and *using control of variables strategy* need further testing also regarding students' understanding of the items. For a more thorough investigation into the effects of the redesigned laboratory on preservice chemistry teacher students' objectively measured and self-assessed SR skills, a main study will be conducted. It remains to be seen what impact the interventions will have on students' scientific reasoning.

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

Funding: This research was funded by (a) Project PEGASUS: Niedersächsisches Ministerium für Wissenschaft und Kultur as part of their program "Innovative Lehr- und Lernkonzepte Innovation Plus" (2019/20), grant number 93 (b) DiBS Chemie: The project is part of the "Qualitätsoffensive Lehrerbildung", a joint initiative of the Federal Government and the Länder which aims to improve the quality of teacher training. The program is funded by the Federal Ministry of Education and Research, grant number 01JA2028. The authors are responsible for the content of this publication. Funding of APC: The authors acknowledge support by the Open Access Publication Funds of Technische Universität Braunschweig.

Institutional Review Board Statement: All participants were students at a German university. They took part voluntarily and signed an informed consent form. Video data were collected with cameras positioned to only film participants' hands and working surfaces. Students who did not wish to be videotaped were seated at a table without recording equipment. Pseudonymization of participants was guaranteed during the study. Due to all these measures in the implementation of the study, an audit by an ethics committee was waived.

Informed Consent Statement: Written informed consent was obtained from all participants in the study.

Data Availability Statement: Video data are not publicly available due to privacy reasons. Other data is available upon request from the authors.

Acknowledgments: We thank Lisanne Kraeva for critical remarks on the adapted instrument and support with data collection, Jaqueline Jacob and Kristina Schaate for support on video production, Dominik Stockmann and Kristina Schaate for support in transcription, data input and rating as well as K. Sachse and all student research assistants for support in the laboratory. Moreover, we thank all students of Organische Chemie 0 for their participation in the study. Finally, we are thankful for critical remarks to an earlier draft by the special issue editors and three anonymous reviewers whose recommendations substantially improved the quality of the article.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A Excerpts from Test Booklet and Coding Manual

Task 5

Experimental task

*The liquids to be compared are water, ethanol and cooking oil. The aim is to determine which of the liquids has a higher surface tension.
(Note: The greater the surface tension of a liquid, the more voluminous drops it can form).*

a) **Generate a hypothesis** about the ranking of liquids according to increasing surface tension.

b) **Plan an experiment** to verify your assumption. Describe exactly how you proceeded! Can you **draw a conclusion**? Give reasons!

Material: Water, ethanol, cooking oil, coins, graduated pipettes, Peleus ball, test tubes (large and small), paper ruler, beakers, Pasteur pipettes, protective goggles

Task 5

Task 5

Except for the studied parameters, there are other conditions for strong surface tension.

a) **Develop a further research question** on this.

b) **Generate a hypothesis** to your research question.

c) **Plan an experiment** to test your hypothesis.

(a)
(b)

Figure A1. (a) The experimental task “surface tension,” measuring SR skills generating hypotheses I, planning experiments I, and drawing conclusions; (b) Task 5, measuring SR skills developing research questions, generating hypotheses II, and planning experiments II.

Task 6

Task 6

The so-called pendant drop method can be used to measure the surface tension of liquids.

In this method, a drop of the liquid is created at the end of a thin tube, e.g. a Pasteur pipette, and its size is measured.

From this, the surface tension is calculated.

The higher the surface tension, the larger the drop can become on the pipette.

liquid	Surface tension in mN/m
Ethanol	23
Water	72,8
Olive oil	32
Glycerine	64,4
Toluene	28,2

Present the measured data in the form of a diagram.

Figure A2. Task 6. Preparing data for analysis.

Table A1. Excerpt from coding manual for SR skills developing research questions, generating hypotheses I/II, planning experiments I/II, observing and measuring, preparing data for analysis, drawing conclusions. A maximum score of 3 points can be achieved in each skill. Based on [26–28,90].

Developing Research Questions	Points
No question is formulated or the question does not address the topic.	0
The question addresses the subject and can be answered using scientific methods.	1
The question is formulated intelligibly and as an open-ended question.	1
The variables specified in the question denote general concepts (not individual cases) ¹ .	1
Generating Hypotheses I/II	Points
No hypothesis is generated or the hypothesis does not address the topic and/or the statement is formulated using “may,” “might,” “could,” “can,” or other expressions differentiating a scientific hypothesis from a mere assumption.	0
A prediction or hypothesis addressing the topic is formulated.	1
The prediction/hypothesis is complemented by an explanation in one or more sentences. The guess/hypothesis is investigable. The guess/hypothesis is falsifiable.	1
The prediction/hypothesis specifies a relationship between to variables (can also be represented by bullet points or arrows/drawings).	1
Planning Experiments I/II	Points
No experiments are named or planned.	0
The student explicates planning (also partial steps).	1
The student plans (and executes ²) an experiment that is not suitable.	2
The student plans (and executes ²) an experiment that is suitable.	3
Observing and Measuring	Points
No observation or measurement is explicated or the observation/measurement is entirely incorrect or the observation/measurement does not address the topic.	0
The student explicates that he/she is observing/measuring. The observation/measurement is relevant to the topic and refers to what is happening in the experiment. Few mistakes are made in the (order of the) observation/measurement.	1
The observation/measurement contains the essential elements of what is happening in the experiment. Data are recorded correctly but using an unsuitable method of measurement.	2
The observation/measurement is purposeful, exhaustive and correct. Data are recorded correctly by using a suitable method of measurement.	3
Preparing Data for Analysis	Points
Task is not answered.	0
Correct type of diagram (line graph/bar chart) is chosen. Variables are correctly assigned to the axes. Axes labels (arrows, categories, or physical quantities and respective units of measurement) are correct.	1
Ratio scales start at zero or explicitly show that the range does not start at zero. Similar distances on a ratio scale denote similar differences in the physical quantity. Tick mark labels/category labels are provided. Lengths of the axes are chosen sensibly.	1
The diagram is neatly drawn. All data points/bars are plotted. Diagram does not extend beyond the specified drawing area. Data points/bars are legible and displayed uniformly and neatly.	1
Drawing Conclusions	Points
Task is not answered or answer does not address the topic.	0
The student names a result.	1
The student’s answer is related to the hypothesis (confirmation or rejection).	1
The student’s answer is based on his/her observation/measurement.	1

¹ Anchor example: general concept: “surface tension of liquids”; individual case: “surface tension of water”. ² Only relevant for rating of SR skill “planning experiments I” in videotaped sequence.

References

- Zimmerman, C.; Klahr, D. Development of Scientific Thinking. In *Stevens’ Handbook of Experimental Psychology and Cognitive Neuroscience*, 4th ed.; Wixted, J.T., Ed.; John Wiley & Sons Inc.: New York, NY, USA, 2018; pp. 1–25.
- Zimmerman, C. The Development of Scientific Reasoning Skills. *Dev. Rev.* **2000**, *20*, 99–149. [CrossRef]
- Piaget, J. The Stages of the Intellectual Development of the Child. In *Educational Psychology in Context. Readings for Future Teachers*; Marlowe, B.A., Canestrari, A.S., Eds.; Sage Publications: Thousand Oaks, CA, USA, 2006; pp. 98–106.
- Lederman, N.G.; Niess, M.L. Problem Solving and Solving Problems: Inquiry About Inquiry. *Sch. Sci. Math.* **2000**, *100*, 113–116. [CrossRef]
- Windschitl, M. Inquiry projects in science teacher education. What can investigative experiences reveal about teacher thinking and eventual classroom practice? *Sci. Educ.* **2003**, *87*, 112–143. [CrossRef]

6. Banchi, H.; Bell, R. The Many Levels of Inquiry. *Sci. Child.* **2008**, *46*, 26–29.
7. Constantinou, C.P.; Tsivitanidou, O.E.; Rybska, E. What Is Inquiry-Based Science Teaching and Learning? In *Professional Development for Inquiry-Based Science Teaching and Learning*; Tsivitanidou, O.E., Ed.; Springer: Cham, Switzerland, 2018; Volume 5, pp. 1–23.
8. Shavelson, R.J. On the measurement of competency. *Empir. Res. Vocat. Educ. Train.* **2010**, *2*, 41–63. [CrossRef]
9. Shavelson, R.J. Assessing business-planning competence using the Collegiate Learning Assessment as a prototype. *Empir. Res. Vocat. Educ. Train.* **2012**, *4*, 77–90. [CrossRef]
10. Hartmann, S.; Upmeyer von Belzen, A.; Krüger, D.; Pant, H.A. Scientific Reasoning in Higher Education. *Z. Psychol.* **2015**, *223*, 47–53. [CrossRef]
11. American Association for the Advancement of Science. *Atlas of Science Literacy Volume 1*; American Association for the Advancement of Science: Washington, DC, USA, 2007.
12. Department for Education. Science Programmes of Study: Key Stage 4. 2014. Available online: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/381380/Science_KS4_PoS_7_November_2014.pdf (accessed on 15 August 2016).
13. Kultusministerkonferenz. *Bildungsstandards Im Fach Chemie Für Den Mittleren Schulabschluss. Beschluss Der Kultusministerkonferenz Vom 16.12.2004*; Luchterhand: München, Germany, 2005. Available online: http://db2.nibis.de/1db/cuvo/datei/bs_ms_kmk_chemie.pdf (accessed on 27 March 2019).
14. Kultusministerkonferenz. *Bildungsstandards Im Fach Chemie Für Die Allgemeine Hochschulreife. Beschluss Der Kultusministerkonferenz Vom 18.06.2020*; Wolters Kluwer: Bonn, Germany, 2020. Available online: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2020/2020_06_18-BildungsstandardsAHR_Chemie.pdf (accessed on 27 April 2021).
15. Morrell, P.D.; Park Rogers, M.A.; Pyle, E.J.; Roehrig, G.; Veal, W.R. Preparing Teachers of Science for 2020 and Beyond. Highlighting Changes to the NSTA/ASTE Standards for Science Teacher Preparation. *J. Sci. Teach. Educ.* **2020**, *31*, 1–7. [CrossRef]
16. Organisation for Economic Co-operation and Development. *PISA 2006 Science Competencies for Tomorrow's World. Volume 1: Analysis*; OECD: Paris, France, 2007. [CrossRef]
17. National Research Council. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*; The National Academies Press: Washington, DC, USA, 2012. [CrossRef]
18. Hacıömeroğlu, E.S.; Hacıömeroğlu, G. Öğretmen adaylarının mantıksal düşünme becerilerinin incelenmesi: Longeot bilişsel gelişim testi. *Türk Bilgisayar ve Matematik Eğitimi Dergisi* **2018**, *9*, 413–448.
19. Milli Eğitim Bakanlığı. Ortaöğretim Kimya Dersi öğretim Programı. (9, 10, 11 ve 12. sınıflar). 2018. Available online: <https://mufredat.meb.gov.tr/Dosyalar/201812102955190-19.01.2018%20Kimya%20Dersi%20%C3%96%C4%9Fretim%20Program%C4%B1.pdf> (accessed on 18 August 2021).
20. Bell, R.L.; Smetana, L.; Binns, I. Simplifying Inquiry Instruction. Assessing the Inquiry Level of Classroom Activities. *Sci. Teach.* **2005**, *72*, 30–34.
21. Cheung, D. Facilitating Chemistry Teachers to Implement Inquiry-based Laboratory Work. *Int. J. Sci. Math. Educ.* **2007**, *6*, 107–130. [CrossRef]
22. Capps, D.K.; Crawford, B.A.; Constan, M.A. A Review of Empirical Literature on Inquiry Professional Development. Alignment with Best Practices and a Critique of the Findings. *J. Sci. Teach. Educ.* **2012**, *23*, 291–318. [CrossRef]
23. Kultusministerkonferenz. *Ländergemeinsame Inhaltliche Anforderungen Für Die Fachwissenschaften und Fachdidaktiken in Der Lehrerbildung. Beschluss der Kultusministerkonferenz Vom 16.10.2008 i. d. F. vom 16.05.2019*. Available online: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf (accessed on 14 July 2017).
24. Mayer, J. Erkenntnisgewinnung als wissenschaftliches Problemlösen. In *Theorien in der Biologiedidaktischen Forschung*; Krüger, D., Vogt, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 177–186.
25. Opitz, A.; Heene, M.; Fischer, F. Measuring scientific reasoning—A review of test instruments. *Educ. Res. Eval.* **2017**, *23*, 78–101. [CrossRef]
26. Kambach, M. *Experimentierprozesse von Lehramtsstudierenden der Biologie*; Logos-Verlag: Berlin, Germany, 2018.
27. Kraeva, L. *Problemlösestrategien von Schülerinnen und Schülern diagnostizieren*; Logos-Verlag: Berlin, Germany, 2020.
28. Nawrath, D.; Maiseyken, V.; Schecker, H. Experimentelle Kompetenz. Ein Modell für die Unterrichtspraxis. *Prax. Nat. Phys. Sch.* **2011**, *60*, 42–49.
29. Klahr, D.; Dunbar, K. Dual Space Search During Scientific Reasoning. *Cogn. Sci.* **1988**, *12*, 1–48. [CrossRef]
30. Fischer, F.; Kollar, I.; Ufer, S.; Sodian, B.; Hussmann, H.; Pekrun, R.; Neuhaus, B.; Dorner, B.; Pankofer, S.; Fischer, M.; et al. Scientific Reasoning and Argumentation: Advancing an Interdisciplinary Research Agenda in Education. *Frontline Learn. Res.* **2014**, *5*, 28–45. [CrossRef]
31. Kambeyo, L. The Possibilities of Assessing Students' Scientific Inquiry Skills Abilities Using an Online Instrument. A Small-Scale Study in the Omusati Region, Namibia. *Eur. J. Educ. Sci.* **2017**, *4*, 1–21. [CrossRef]
32. Yanto, B.E.; Subali, B.; Suyanto, S. Improving Students' Scientific Reasoning Skills through the Three Levels of Inquiry. *Int. J. Instr.* **2019**, *12*, 689–704. [CrossRef]
33. Wellnitz, N.; Fischer, H.E.; Kauertz, A.; Mayer, J.; Neumann, I.; Pant, H.A.; Sumfleth, E.; Walpuski, M. Evaluation der Bildungsstandards—eine fächerübergreifende Testkonzeption für den Kompetenzbereich Erkenntnisgewinnung. *Z. Didakt. Nat.* **2012**, *18*, 261–291.

34. Krell, M.; Redman, C.; Mathesius, S.; Krüger, D.; van Driel, J. Assessing Pre-Service Science Teachers' Scientific Reasoning Competencies. *Res. Sci. Educ.* **2020**, *50*, 2305–2329. [CrossRef]
35. Chang, H.-P.; Chen, C.-C.; Guo, G.-J.; Cheng, Y.-J.; Lin, C.-Y.; Jen, T.-H. The Development of a Competence Scale for Learning Science. Inquiry and Communication. *Int. J. Sci. Math. Educ.* **2011**, *9*, 1213–1233. [CrossRef]
36. Kunz, H. Professionswissen von Lehrkräften der Naturwissenschaften im Kompetenzbereich Erkenntnisgewinnung. Ph.D. Thesis, Universität Kassel, Kassel, Germany, 2011. Available online: <https://kobra.bibliothek.uni-kassel.de/bitstream/urn:nbn:de:hebis:34-2012012040403/9/DissertationHagenKunz.pdf> (accessed on 9 August 2021).
37. Grube, C.R. Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung. Untersuchung der Struktur und Entwicklung des wissenschaftlichen Denkens bei Schülerinnen und Schülern der Sekundarstufe, I. Ph.D. Thesis, Universität Kassel, Kassel, Germany, 2011. Available online: <https://kobra.uni-kassel.de/handle/123456789/2011041537247> (accessed on 21 August 2021).
38. Schwichow, M.G.; Croker, S.; Zimmerman, C.; Höffler, T.N.; Härtig, H. Teaching the control-of-variables strategy. A meta-analysis. *Dev. Res.* **2016**, *39*, 37–63. [CrossRef]
39. Renkl, A.; Mandl, H.; Gruber, H. Inert Knowledge: Analyses and Remedies. *Educ. Psychol.* **1996**, *31*, 115–121. [CrossRef]
40. Leuders, T.; Naccarella, D.; Philipp, K. Experimentelles Denken—Vorgehensweisen beim innermathematischen Experimentieren. *J. Math.-Didakt.* **2011**, *32*, 205–231. [CrossRef]
41. Okada, T.; Simon, H.A. Collaborative Discovery in a Scientific Domain. *Cogn. Sci.* **1997**, *21*, 109–146. [CrossRef]
42. Khan, S.; Krell, M. Scientific Reasoning Competencies: A Case of Preservice Teacher Education. *Can. J. Sci. Math. Technol. Educ.* **2019**, *19*, 446–464. [CrossRef]
43. Hilfert-Rüppell, D.; Looß, M.; Klingenberg, K.; Eghtessad, A.; Höner, K.; Müller, R.; Strahl, A.; Pietzner, V. Scientific reasoning of prospective science teachers in designing a biological experiment. *Lehr. Auf. Dem. Prüfstand* **2013**, *6*, 135–154.
44. Hofstein, A.; Lunetta, V.N. The Laboratory in Science Education. Foundations for the Twenty-First Century. *Sci. Educ.* **2004**, *88*, 28–54. [CrossRef]
45. Koenig, K.; Schen, M.; Bao, L. Explicitly Targeting Pre-service Teacher Scientific Reasoning Abilities and Understanding of Nature of Science through an Introductory Science Course. *Sci. Educ.* **2012**, *21*, 1–9.
46. Bruckermann, T.; Schlüter, K. *Forschendes Lernen im Experimentalpraktikum Biologie*; Springer: Berlin/Heidelberg, Germany, 2017. [CrossRef]
47. Fischer, R.A. Den Pulsschlag der Chemie fühlen—schon im Grundpraktikum. *Angew. Chem.* **2017**, *129*, 7792–7793. [CrossRef]
48. Johnstone, A.H. Chemistry Teaching—Science or Alchemy? *J. Chem. Educ.* **1997**, *74*, 262. [CrossRef]
49. Kirschner, P.A.; Sweller, J.; Clark, R.E. Why Minimal Guidance During Instruction Does Not Work. An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educ. Psychol.* **2006**, *41*, 75–86. [CrossRef]
50. Hmelo-Silver, C.E.; Duncan, R.G.; Chinn, C.A. Scaffolding and Achievement in Problem-Based and Inquiry Learning. A Response to Kirschner, Sweller, and Clark (2006). *Educ. Psychol.* **2007**, *42*, 99–107. [CrossRef]
51. Mulder, Y.G.; Lazonder, A.W.; de Jong, T. Using heuristic worked examples to promote inquiry-based learning. *Learn. Instr.* **2014**, *29*, 56–64. [CrossRef]
52. Tuovinen, J.E.; Sweller, J. A comparison of cognitive load associated with discovery learning and worked examples. *J. Educ. Psychol.* **1999**, *91*, 334–341. [CrossRef]
53. Leppink, J.; Paas, F.; van Gog, T.; van der Vleuten, C.P.M.; van Merriënboer, J.J.G. Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learn. Instr.* **2014**, *30*, 32–42. [CrossRef]
54. Singh, C. Interactive video tutorials for enhancing problem-solving, reasoning, and meta-cognitive skills of introductory physics students. *AIP Conf. Proc.* **2004**, *720*, 177–180. [CrossRef]
55. Lazonder, A.W.; Kamp, E. Bit by bit or all at once? Splitting up the inquiry task to promote children's scientific reasoning. *Learn. Instr.* **2012**, *22*, 458–464. [CrossRef]
56. Lazonder, A.W.; Harmsen, R. Meta-Analysis of Inquiry-Based Learning. *Rev. Educ. Res.* **2016**, *86*, 681–718. [CrossRef]
57. Lazonder, A.W.; Hagemans, M.G.; de Jong, T. Offering and discovering domain information in simulation-based inquiry learning. *Learn. Instr.* **2010**, *20*, 511–520. [CrossRef]
58. Kaiser, I.; Mayer, J. The Long-Term Benefit of Video Modeling Examples for Guided Inquiry. *Front. Educ.* **2019**, *4*, 1–18. [CrossRef]
59. Kant, J.M. Fostering the Acquisition of Scientific Reasoning with Video Modeling Examples and Inquiry Tasks. Ph.D. Thesis, Eberhard Karls Universität Tübingen, Tübingen, Germany, 2017. Available online: <https://publikationen.uni-tuebingen.de/xmlui/handle/10900/74159> (accessed on 18 August 2021).
60. Ramachandran, R.; Sparck, E.M.; Levis-Fitzgerald, M. Investigating the Effectiveness of Using Application-Based Science Education Videos in a General Chemistry Lecture Course. *J. Chem. Educ.* **2019**, *96*, 479–485. [CrossRef]
61. Pulukuri, S.; Abrams, B. Improving Learning Outcomes and Metacognitive Monitoring. Replacing Traditional Textbook Readings with Question-Embedded Videos. *J. Chem. Educ.* **2021**, *98*, 2156–2166. [CrossRef]
62. Jolley, D.F.; Wilson, S.R.; Kelso, C.; O'Brien, G.; Mason, C.E. Analytical Thinking, Analytical Action. Using Prelab Video Demonstrations and e-Quizzes to Improve Undergraduate Preparedness for Analytical Chemistry Practical Classes. *J. Chem. Educ.* **2016**, *93*, 1855–1862. [CrossRef]
63. Lewis, R.A. Video introductions to laboratory. Students positive, grades unchanged. *Am. J. Phys.* **1995**, *63*, 468–470. [CrossRef]
64. Campbell, J.; Macey, A.; Chen, W.; Shah, U.V.; Brechtelsbauer, C. Creating a Confident and Curious Cohort. The Effect of Video-Led Instructions on Teaching First-Year Chemical Engineering Laboratories. *J. Chem. Educ.* **2020**, *97*, 4001–4007. [CrossRef]

65. Seery, M.K.; Agustian, H.Y.; Doidge, E.D.; Kucharski, M.M.; O'Connor, H.M.; Price, A. Developing laboratory skills by incorporating peer-review and digital badges. *Chem. Educ. Res. Pract.* **2017**, *18*, 403–419. [CrossRef]
66. Stieff, M.; Werner, S.M.; Fink, B.; Meador, D. Online Prelaboratory Videos Improve Student Performance in the General Chemistry Laboratory. *J. Chem. Educ.* **2018**, *95*, 1260–1266. [CrossRef]
67. Brame, C.J. Effective Educational Videos. Principles and Guidelines for Maximizing Student Learning from Video Content. *CBE Life Sci. Educ.* **2016**, *15*, 1–6. [CrossRef]
68. Schnotz, W.; Kürschner, C.A. Reconsideration of Cognitive Load Theory. *Educ. Psychol. Rev.* **2007**, *19*, 469–508. [CrossRef]
69. Mayer, R.E. (Ed.) *The Cambridge Handbook of Multimedia Learning*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2014.
70. Bravo, E.; Amante, B.; Simo, P.; Enache, M.; Fernandez, V. Video as a new teaching tool to increase student motivation. In Proceedings of the IEEE Global Engineering Education Conference (EDUCON), Amman, Jordan, 4–6 April 2011; pp. 638–642.
71. Harwood, W.S.; McMahan, M.M. Effects of Integrated Video Media on Student Achievement and Attitudes in High School Chemistry. *J. Res. Sci. Teach.* **1997**, *34*, 617–631. [CrossRef]
72. Jenö, L.M.; Vandvik, V.; Eliassen, S.; Grytnes, J.-A. Testing the novelty effect of an m-learning tool on internalization and achievement. A Self-Determination Theory approach. *Comput. Educ.* **2019**, *128*, 398–413. [CrossRef]
73. Huang, W. Investigating the Novelty Effect in Virtual Reality on STEM Learning. Ph.D. Thesis, Arizona State University, Tempe, AZ, USA, 2020. Available online: https://repository.asu.edu/attachments/227504/content/huang_asu_0010E_20075.pdf (accessed on 18 August 2021).
74. Hansford, B.C.; Hattie, J.A. The Relationship between Self and Achievement/Performance Measures. *Rev. Educ. Res.* **1982**, *52*, 123. [CrossRef]
75. Marsh, H.W.; Craven, R.G. Reciprocal Effects of Self-Concept and Performance from a Multidimensional Perspective. Beyond Seductive Pleasure and Unidimensional Perspectives. *Perspect. Psychol. Sci.* **2006**, *1*, 133–163. [CrossRef] [PubMed]
76. House, D.J. Noncognitive Predictors of Achievement in Introductory College Chemistry. *Res. High. Educ.* **1995**, *36*, 473–490. [CrossRef]
77. Atzert, R.; John, R.; Preisfeld, A.; Damerau, K. Der Einfluss kriterialer, sozialer und individueller Bezugsnormen auf das experimentbezogene Fähigkeitsselbstkonzept. *Z. Didakt. Nat.* **2020**, *26*, 89–102. [CrossRef]
78. Sudria, I.B.N.; Redhana, I.W.; Suja, I.W.; Suardana, I.N. Self-assessment of chemistry laboratory basic skills using performance scoring rubrics at the chemistry teacher training. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *959*, 1–8. [CrossRef]
79. Schöne, C.; Dickhäuser, O.; Spinath, B. Das Fähigkeitsselbstkonzept und seine Erfassung. In *Diagnostik von Motivation und Selbstkonzept*; Stiensmeier-Pelster, J., Rheinberg, F., Eds.; Hogrefe: Göttingen, Germany, 2002; pp. 3–14.
80. Technische Universität Braunschweig Fakultät 6. *Besonderer Teil der Prüfungsordnung für den Bachelorstudiengang Erziehungswissenschaft und den Zwei-Fächer-Bachelorstudiengang der der Technischen Universität Braunschweig inkl. der 8. Änderung*; Nichtamtliche Lesefassung; Präsident der Technischen Universität Braunschweig, Ed.; Technische Universität Braunschweig: Braunschweig, Germany, 2013. Available online: <https://www.tu-braunschweig.de/index.php?eID=dumpFile&t=f&f=87690&token=ea5059c4421ac247a58e9a6031970b9c648b30bd> (accessed on 12 November 2020).
81. Bicak, B.E.; Borchert, C.E.; Höner, K. *Strategy to implement inquiry instructions in an organic chemistry lab course*; TU Braunschweig: Braunschweig, Germany, 2021; Manuscript in preparation.
82. Christian, G.D. *Analytical Chemistry*, 4th ed.; John Wiley & Sons: New York, NY, USA, 1986.
83. Danzer, K. *Analytical Chemistry Theoretical and Meteorological Fundamentals*; Springer: Berlin/Heidelberg, Germany, 2007.
84. Hardcastle, W.A. *Qualitative Analysis: A Guide to Best Practice*; Royal Society of Chemistry: London, UK, 1998.
85. Bicak, B.E.; Borchert, C.; Höner, K. Förderung von Erkenntnisgewinnung mit experimentellem Problemlösen und Lernvideos im organisch-chemischen Praktikum. In *Naturwissenschaftlicher Unterricht und Lehrerbildung im Umbruch*; Habig, S., Ed.; GDGP: Essen, Germany, 2021; Volume 41, pp. 334–337. Available online: https://www.gdgp-ev.de/wp-content/tb2021/TB2021_334_Bicak.pdf (accessed on 22 July 2021).
86. Metzger, S.; Sommer, K. “Kochrezept” oder experimentelle Methode? *MNU J.* **2010**, *68*, 4–11.
87. Baur, A.; Hummel, E.; Emden, M.; Schröter, E. Wie offen sollte offenes Experimentieren sein? Ein Plädoyer für das geöffnete Experimentieren. *MNU J.* **2020**, *73*, 125–128.
88. Mackensen-Friedrichs, I. Förderung des Expertiserwerbs durch das Lernen mit Beispielaufgaben im Biologieunterricht der Klasse 9. Ph.D. Dissertation, Christian-Albrechts-Universität Kiel, Kiel, Germany, 2004.
89. Arnold, J.C.; Kremer, K.; Mayer, J. Understanding Students’ Experiments—What kind of support do they need in inquiry tasks? *Int. J. Sci. Educ.* **2004**, *36*, 2719–2749. [CrossRef]
90. Lachmayer, S. Entwicklung und Überprüfung eines Strukturmodells der Diagrammkompetenz für den Biologieunterricht. Ph.D. Dissertation, Christian-Albrechts-Universität, Kiel, Germany, 2008.
91. Technische Universität Braunschweig. Die Technische Universität Braunschweig in Zahlen 2020. *Technische Universität Braunschweig*. Available online: https://www.tu-braunschweig.de/fileadmin/Redaktionsgruppen/Stabsstellen/SPK/ordnungen-leitlinien-fakten/tubraunschweig_zahlen.pdf (accessed on 12 November 2020).
92. American Psychological Association. Ceiling Effect. APA Dictionary of Psychology. 2020. Available online: <https://dictionary.apa.org/ceiling-effect> (accessed on 18 August 2021).
93. Bhattacharjee, A. *Social Science Research. Principles, Methods, and Practices*, 2nd ed.; University of South Florida Scholar Commons: Tampa, FL, USA, 2012. Available online: https://digitalcommons.usf.edu/oa_textbooks/3/ (accessed on 18 August 2021).

94. Schmitt, N. Uses and Abuses of Coefficient Alpha. *Psychol. Assess.* **1996**, *8*, 350–353. [CrossRef]
95. Schecker, H. Überprüfung der Konsistenz von Itemgruppen mit Cronbachs α . In *Methoden in der Naturwissenschaftsdidaktischen Forschung*; Krüger, D., Parchmann, I., Schecker, H., Eds.; Springer-Spektrum: Berlin, Germany, 2014. Available online: <https://static.springer.com/sgw/documents/1426184/application/pdf/Cronbach+Alpha.pdf> (accessed on 30 November 2020).
96. Cohen, J. A power primer. *Psychol. Bull.* **1992**, *112*, 155–159. [CrossRef]
97. Liljequist, D.; Elfving, B.; Skavberg Roaldsen, K. Intraclass correlation—A discussion and demonstration of basic features. *PLoS ONE* **2019**, *14*, e0219854. [CrossRef]
98. Cicchetti, D.V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol. Assess.* **1994**, *6*, 284–290. [CrossRef]
99. Csanadi, A.; Kollar, I.; Fischer, F. Pre-service teachers' evidence-based reasoning during pedagogical problem-solving: Better together? *Eur. J. Psychol. Educ.* **2021**, *36*, 147–168. [CrossRef]
100. Stiller, J.; Hartmann, S.; Mathesius, S.; Straube, P.; Tiemann, R.; Nordmeier, V.; Krüger, D.; Upmeyer zu Belzen, A. Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assess. Eval. High. Educ.* **2016**, *41*, 721–732. [CrossRef]
101. Krell, M.; Khan, S.; van Driel, J. Analyzing Cognitive Demands of a Scientific Reasoning Test Using the Linear Logistic Test Model (LLTM). *Educ. Sci.* **2021**, *11*, 472. [CrossRef]
102. Bernholt, S.; Parchmann, I. Assessing the complexity of students' knowledge in chemistry. *Chem. Educ. Res. Pract.* **2011**, *12*, 167–173. [CrossRef]

Article

Professional Knowledge and Self-Efficacy Expectations of Pre-Service Teachers Regarding Scientific Reasoning and Diagnostics

Dagmar Hilfert-Rüppell ^{1,*} , Monique Meier ², Daniel Horn ²  and Kerstin Höner ¹

¹ Institut for Science Education, Technische Universität Braunschweig, 38106 Braunschweig, Germany; k.hoener@tu-braunschweig.de

² Biology Education, Universität Kassel, 34132 Kassel, Germany; mmeier@uni-kassel.de (M.M.); daniel.horn@uni-kassel.de (D.H.)

* Correspondence: d.hilfert-ruempel@tu-braunschweig.de

Abstract: Understanding and knowledge of scientific reasoning skills is a key ability of pre-service teachers. In a written survey (open response format), biology and chemistry pre-service teachers ($n = 51$) from two German universities claimed central decisions or actions school students have to perform in scientific reasoning in the open inquiry instruction of an experiment. The participants' answers were assessed in a quality content analysis using a rubric system generated from a theoretical background. Instruments in a closed response format were used to measure attitudes towards the importance of diagnostics in teacher training and the domain-specific expectations of self-efficacy. The pre-service teacher lacked pedagogical (didactics) content knowledge about potential student difficulties and also exhibited a low level of content methodological (procedural) knowledge. There was no correlation between the knowledge of student difficulties and the approach to experimenting with expectations of self-efficacy for diagnosing student abilities regarding scientific reasoning. Self-efficacy expectations concerning their own abilities to successfully cope with general and experimental diagnostic activities were significantly lower than the attitude towards the importance of diagnostics in teacher training. The results are discussed with regard to practical implications as they imply that scientific reasoning should be promoted in university courses, emphasising the importance of understanding the science-specific procedures (knowing how) and epistemic constructs in scientific reasoning (knowing why).

Citation: Hilfert-Rüppell, D.; Meier, M.; Horn, D.; Höner, K. Professional Knowledge and Self-Efficacy Expectations of Pre-Service Teachers Regarding Scientific Reasoning and Diagnostics. *Educ. Sci.* **2021**, *11*, 629. <https://doi.org/10.3390/educsci11100629>

Academic Editors: Moritz Krell, Andreas Vorholzer and Andreas Nehring

Received: 8 August 2021

Accepted: 6 October 2021

Published: 11 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: professional knowledge; scientific reasoning skills; self-efficacy; students' difficulties; diagnostic competencies

1. Introduction

Inquiry-based teaching is seen as contributing to content, procedural, and epistemic learning goals of science education [1]. Therefore, a basic understanding of the systematic approach to conducting science investigations is required in the competencies of scientific inquiry (e.g., in the US [2] and in the UK [3]). In Germany, these are reported for biology, chemistry, and physics with similarly formulated educational standards [4–6], allowing these competencies to be promoted in a networked manner both vertically within a subject and horizontally across subjects [7]. This applies also to the field of school education and in the training of pre-service teachers. Both areas are in turn directly related, i.e., the knowledge and skills (for scientific inquiry) of the teacher shape the teaching with learning opportunities (for inquiry-based science education) and thus the potential learning success on the part of the students [8]. As a result, at the international and national levels, standards for the teaching profession are also being formulated and science teaching competencies described for teacher education [9–11]. Common to them are the requirements for future teachers to build up science education about and through scientific inquiry as their own

competence, as well as to learn how to teach scientific inquiry in order to be able to promote corresponding competencies in students. For example, (ongoing) teachers have to be competent themselves in designing empirical approaches to test hypotheses, and need knowledge as well as skills in hypothesis-led experimentation [12]. Studies on the development of pre-service science teachers' scientific reasoning competencies show that explicit reflections about scientific reasoning (i.e., learning about science; [13]) contributes more to the development of scientific reasoning competencies than only doing science without reflecting about it [14,15]. In the course of this, knowledge about and understanding of scientific inquiry and scientific reasoning is relevant [16], and can also be used as constructs for the analysis and assessment of learning activities and learning outcomes [17].

1.1. Scientific Reasoning

In many studies, the terms scientific reasoning and scientific thinking are used interchangeably because the boundary between reasoning and thinking is blurred (e.g., [18–20]). Scientific reasoning thus forms the basis for critical thinking and is only one, albeit very significant, aspect in the process of thinking about (scientific) facts (e.g., [21–24]). Scientific reasoning can therefore be interpreted as a subset of critical thinking skills (cognitive and metacognitive processes and dispositions) that are essential for scientific procedures in the problem-solving process, the evidence of information in scientific disciplines, and the epistemological incorporation of scientific methods and paradigms [25]. There are a range of views on the structure of scientific reasoning and on the number of its components. Two groupings can be distinguished: while one emphasises scientific reasoning as a broad and complex component representing a particular skill, understanding, or competence [26], the other grouping opposes the advocacy of multidimensional theories [27].

Since the 1970s, according to Dunbar and Fugelsang [28], scientific reasoning has also been viewed as a way of solving problems, with great efforts being made to identify strategies that scientists use to solve problems. Competencies required for scientific reasoning are viewed as a complex construct, encompassing both the skills required for scientific problem solving and the ability to reflect on problem solving at a meta-level [29,30]. The scientific discovery process is best conceptualised as involving both reasoning and problem-solving skills, with the ultimate goal of generating, testing, and then evaluating a hypothesis about a causal or categorical relationship based on the results. Both of these skills—strategy development and reasoning processes—require knowledge to identify key features of a problem at hand [31]. Problem solving requires three types of knowledge in a complex construct (knowing that, knowing how, knowing why, cf., [29,32,33]). Explicit procedural knowledge, in turn, addresses the execution level and thus the “knowing how” and practical implementation of actions to solve problems [32,33]. In his structural model of scientific reasoning, Mayer [34] identified, in addition to personal variables, such as (prior) knowledge and cognition, four process-related skills/subcompetencies:

1. Formulating scientific questions;
2. Generating hypotheses;
3. Planning scientific investigations;
4. Interpreting data, which are differentiated via different aspects of competence, e.g., [35].

In the execution of these process steps in problem-oriented and inquiry-based teaching, content knowledge is to be applied and methodological knowledge is to be developed and applied in equal measure [19]. Experimentation is considered to play a central role in the process of scientific inquiry as a content and method in science education [36]. “Scientific reasoning is defined as the inquiry processes [. . .] [and] the reasoning skills involved in experimentation, evidence evaluation, and inference making addressed to scientific understanding” [37] (p. 106). In the present study, scientific reasoning is predominantly defined from a procedural perspective and thus addressed in the most open-ended problem-solving process possible using the method of experimentation.

In order to perform diagnostics and scaffolding in the (experimental) inquiry process, knowledge about scientific reasoning, as well as about the implementation and associated

difficulties on the part of the learners of scientific reasoning, plays a central role for (pre-service) teachers. Given this significance, the key difficulties faced by students will be outlined below in a summary of the literature.

1.2. Literature Summary on Student Difficulties in the Experimental Problem Solving Process

Experimentation demands and promotes a wide range of cognitive, psychomotor, and social skills in students [35]. In this context, the SDDS model (Scientific Discovery as Dual Search; [22]) as well as the structural model for scientific reasoning [34] with their anchoring in problem-solving research is relevant for both the conceptualisation and the assessment of competencies in the science subjects of biology [37–41], chemistry [42–44], and physics [45–47], respectively, and is central in natural science studies [7,48]. In promoting the competencies underlying each model, the inquiry-based learning approach is shown to be superior to direct instruction on these [49]. In inquiry-based learning, the degree of student activity or, respectively, the open-endedness in the experimentation can be designed differently (levels of inquiry [50]) and thus influence learning success. In the literature, on the one hand, the high value of independence in experimentation compared to teacher demonstration experiments is emphasised (e.g., [51,52]) and, on the other hand, the guided inquiry approach with targeted support of learners in the phases for experimental problem solving is attributed the highest effectiveness [53]. Vorholzer and von Aufschnaiter [54] identify three main dimensions in which the implementation of guidance can vary: (a) the degree of autonomy, (b) the degree of conceptual information, and (c) the cognitive domain of guidance. Independent experimentation can be understood as a relatively complex cognitive problem-solving process that is particularly challenging for students (cf., [55,56]) and consequently requires scaffolding in the different dimensions, process steps, and/or competencies. Accordingly, pre-service science teachers must be provided with learning opportunities to acquire these competencies, to foster their own reasoning in science alongside how to teach students how to reason. They should also be enabled to plan and implement targeted lessons that enable their students to acquire the scientific reasoning competencies for experimentation and experimental problem solving required by the educational standards of many countries [4–6,57]. Neither mere boilerplate imitation of experiments nor participation in teacher demonstration experiments by students leads to the desired mastery of the scientific reasoning process [58]; nor are approaches opened too early effective [59,60]. The latter can create numerous difficulties/barriers for students to overcome during experimentation. According to the structural areas/phases of hypothesis, planning/execution, and testing, as well as conclusions and evaluation of results in the models of Klahr [22] and Mayer [34], the misconceptions and student difficulties described here descriptively and proven empirically are summarised in Table 1.

Table 1. Overview of process-related misconceptions and difficulties of students in experimentation published in the literature (references in brackets on the right side) structured and summarised according to the three phases for scientific reasoning of the SDDS model [22].

Phase:	Search Hypothesis	[39,61–67]
	No hypothesis is formulated.	
	No alternative hypotheses are formulated.	
	There is no idea about the purpose of hypothesising.	
	Hypotheses are not related to the research question.	
	Hypotheses are formulated and/or changed in the scientific inquiry process.	
	In the hypothesis, the variables are not defined or are defined incorrectly.	
	Hypotheses are formulated without justification (conjectures).	
	Assumptions are justified by themselves.	
	Assumptions are justified with reference to everyday life.	

Table 1. Cont.

Phase:	Planning and Testing	[39,56,63,66–76]
	Experiment is not suitable for testing the hypothesis. No plan and/or unstructured trial and error (no plan, change all). Missing and/or incorrect operationalisation of variables. Unsystematic handling of variables, i.e., several variables are confounded with each other or the same variable is varied unsystematically. Planning of an experiment that does not lead to the desired results. Lack of a measurement concept/measurement repetitions. Lack of a control approach and/or attention to control variables. Selection of materials is unsystematic, incomplete and/or done by trial and error. Difficulties in handling (simple) materials (e.g., pipette). Wrong observation focus. Interference factors are perceived but not eliminated.	
Phase:	Evaluate Evidence	[63,66,67,72,75,77–80]
	Hypothesis is confirmed without considering the results (confirmatory bias). Hypothesis is adjusted to the results. Data are not (re)related to hypothesis. Not all results are considered; (unexpected) results are (partly) ignored. Unexpected results are attributed to errors in the experimental procedure. Missing or incomplete (error) reflection of the results. Wrong conclusion from coherent experiments. Due to unsystematic variable variation, no conclusions can actually be drawn, or illogical conclusions are drawn.	

1.2.1. Formulating/Search Hypothesis

When generating hypotheses, students often find it difficult to make educated guesses. They do not consider preconditions such as justification or verifiability—in principle or with the given experimental materials. The procedure that several hypotheses are to be set up is mostly unknown to them or only the one they think is correct is considered (e.g., [64,81]).

1.2.2. Design and Execution of the Experiment

It is often observed that the students do not plan out their approach to the experiment, but instead carry it out immediately and instinctively with no pre-determined method in mind. This can result in the steps in the experiment not being purposeful and the procedure being changed several times [72]. Hammann and colleagues [40] describe this unstructured trial-and-error approach as “no plan, change all”. The most important scientific reasoning skill describes the control of variables strategy [70]. The difficulties of students here lie in identifying the dependent and independent variable ([82,83], among others), and in the fact that the variable control is often not considered and confounding variables are not excluded [63]. Moreover, a control approach is usually missing, and measurement repetitions are rarely performed [39,69,82,83]. Sometimes students try to create an effect rather than conduct a goal-directed experiment [75,78]. Inefficient experimentation is also evident, e.g., the same experiment is repeated multiple times [81]. Kraeva [44] was able to identify six different approaches of students in conducting chemistry experiments through video analysis, classified in terms of the attributes “plan” versus “try” as well as “maintain”, “develop”, and “discard”. “Revision” as a planning-reviewing strategy was described as relatively successful, whereas ‘imitation’ (exploratory-alternative-less) was described as a relatively less successful strategy. This corresponds with the courses of action taken by students when experimenting to clarify a biological phenomenon in the sense of a process-oriented or explorative type [39].

1.2.3. Evaluation of the Evidence

The generated data are partly disregarded, while conclusions are drawn illogically and for the most part not related back to the hypothesis [67]. Hammann ([84], p. 200) refers to the hasty termination of the search for possible hypotheses and thus to the wrong conclusion as “positive capture”. He calls the “most robust finding in the literature” the “confirmation bias” [62,79], i.e., the tendency to confirm hypotheses while ignoring contradictory data [84]. Dunbar [85] distinguishes between two approaches to data evaluation: the “find-evidence-goal” approach, i.e., looking for results that confirm the hypothesis, and the “find-hypothesis-goal” approach, i.e., looking for new hypotheses after non-confirming results that then reflect the results. Evaluation of the evidence in the form of referring back to the hypothesis rarely occurs, as does discussion of error [39].

The probability of success for a specific experimentation process depends on the one hand, on the requirements resulting from the individual phases of the experimentation process [86], and, on the other hand, on personal characteristics, such as interest in the subject or general cognitive performance of the students [55]. For diagnosis, in this case the assessment of students’ abilities and performance in experimental problem solving, the discrepancy between students’ conceptions and students’ actions on scientific concepts for experimentation is used [87,88]. Draude [89] was able to demonstrate deficits regarding the diagnosis of student difficulties in experimentation for physics teachers, whereby the necessary prerequisites are hardly developed and promoted in the teacher training programs [75,90]. In a longitudinal two-year study with biology pre-service teachers, insights into the structure and change of their diagnostic competence and possible influencing factors were obtained [91]. Diagnoses of student difficulties and performance in experimental problem solving can only succeed if pre-service teachers themselves have the appropriate (professional) knowledge of the subject matter to be diagnosed. Otherwise, diagnostic processes will be impaired by the existing knowledge gaps [92]. Before a promotion in this area can and should be targeted and discussed as a training element in teacher education, we pursue with this study the concern to describe pre-service science teachers’ scientific reasoning competencies in order to derive the relevance of possible curricular implementations in subject, subject didactics, and educational science for diagnostics in experimentation. As described, the latter is of importance in all study elements, but is located differently and requires a theory-related consideration in the following.

1.3. Relevance of Diagnostic Competencies for (Pre-Service) Teachers

Diagnostic competence of teachers describes both the ability to successfully cope with the diagnostic tasks arising in the teaching profession and the quality of the diagnostic performance [93]. Making efficient instructionally relevant decisions is impossible without being able to identify, understand, and even predict instructionally relevant situations and events [94]. Thus, the investment of a (subject-related) diagnostic competence of (pre-service) teachers seems to be an indispensable prerequisite for the teaching profession [95]. In Shulman’s three main domains of professional knowledge (content knowledge (CK), pedagogical content knowledge (PCK), pedagogical knowledge (PK) [96], which is the most widely used classification in the literature, knowledge about assessment and diagnosis is classified in the domain of general didactical knowledge (pedagogical knowledge, PK) [97]. Kramer et al. [95] describe either PK (more generic: e.g., teaching disorder [98]) or the subject-specific facets CK or PCK (e.g., diagnosing biology instruction, [99]) as relevant to the application of diagnostic activities and diagnostic accuracy, depending on the diagnostic focus. Results of path analyses utilising Rasch measures showed that both PCK and PK were statistically significantly in relation to pre-service teachers’ diagnostic activities. Additionally, biology teachers’ PCK was positively related to diagnostic accuracy [95].

Divergent assumptions exist about what comprises teachers’ diagnostic competence, stemming from the fact that different aspects such as subject matter, method, and target are modeled. The conceptualisation of diagnostic activities in which knowledge is applied in order to solve specific problems can be seen as equivalent to scientific reasoning

skills [100]. Crucial for a sustainable diagnostic cycle is the transformation of competence into performance mediated by situational skills of perception (P), interpretation (I), and (action) decision (D) in the sense of Blömeke, Gustafsson, and Shavelson's [101] model, in which teachers' competence is viewed as a continuum with multiple transitions (P-I-D model of competence transformation). In this context, diagnostic competence presupposes the correct perception of relevant classroom features (noticing) and their evaluation with reference to theoretically grounded, pedagogical action knowledge (reasoning) [102]. Draude [89] distinguishes between predictive and action diagnostic competence of physics teachers. While the predictive competence measured the extent to which teachers could predict students' experimental difficulties in a particular physics experiment, the action-accompanying diagnostic competence measured the extent to which teachers diagnosed difficulties during the students' experimentation process. He also found deficits in both areas, so that a promotion of (pre-service) teachers' skills in this regard seems to be indicated [89]. In the present study, therefore, the predictive diagnostic competence of biology and chemistry pre-service teachers with regard to student difficulties, misconceptions, and the necessary central decisions students have to make during open-ended experimentation was assessed by means of a text-based description of a teaching scenario for a student experiment. In terms of examining diagnostic competence, self-report is common in research, so there is a need for tools to survey diagnostic and reflective skills in a natural setting [103,104].

1.4. Self-Efficacy Expectations

Self-efficacy expectations are considered another major aspect of teachers' professional competence; following Baumert and Kunter's [105] model of professional competence, self-efficacy expectations are considered relevant in addition to knowledge and attitudes. Self-assessments related to motivation, personal engagement, or self-efficacy also appear to be of value in better understanding the interplay between motivational and affective states and diagnostic activities. Self-efficacy, first introduced by Bandura [106] as an aspect of social cognitive learning theory, is described as the strength of one's belief in one's ability to perform a particular task or achieve a particular outcome. Thus, assessing self-efficacy is less about what skills and abilities individuals possess and more about what they believe they can do with the skills and abilities they possess [107]. In this regard, competent performance is guided in part by higher-level self-regulatory abilities [108]. These include general abilities to diagnose task demands, construct and evaluate alternative courses of action, set perspective-close goals to guide one's efforts, and create self-incentives to maintain engagement in stressful activities and manage stress and distracting thoughts [109]. Self-efficacy correlates with academic performance [110,111], task persistence, motivation [112], and resilience in academic contexts [113]. Self-efficacy varies depending on the situation and therefore needs to be considered or captured in a domain- and context-specific manner [114]. Students' self-efficacy in science education has been studied in the science subjects of mathematics [115–118], physics [117], and chemistry [118,119]. With regard to problem solving in mathematics, it was shown that even when students have the ability to solve problems, those who have a strong self-efficacy expectancy are more effective problem solvers [116,117].

Regarding the self-efficacy expectancy of (pre-service) teachers in science, there are some less empirical findings [120–122]. In a study by Yürük [120], pre-service teachers who had taken more science courses in college, felt better prepared to teach science content and had higher levels of self-efficacy. Riese and Reinhold [123] addressed the relationship between physics teachers' CK and PCK (compare [96], see above) and their general and classroom self-efficacy. They found a significant positive correlation between teaching-related self-efficacy and CK. Kurbanoglu and Akim [121], meanwhile, show that low self-efficacy expectancy regarding the subject of chemistry predicts chemistry laboratory anxiety and has a negative effect on freshmen's attitudes toward chemistry. In biology, there have been very few studies on self-efficacy (compare [114,122,124]). The findings of

Mahler, Großschedl, and Harms [124] indicate that teacher education in college, attending professional development courses, and self-study provide learning opportunities to promote self-efficacy and enthusiasm for teaching. In addition, the authors found that self-efficacy and subject-specific enthusiasm were positively related to PCK.

While self-efficacy or self-efficacy expectations are increasingly a focus of inquiry in teacher education, very few studies can be identified that address the application of self-efficacy theory to diagnostic skills or self-reported perceptions of self-efficacy as a predictor of actual diagnostic skills in pre-service teachers. Motivation, attitude, and knowledge were found to be significant positive predictors of diagnostic skills with respect to learning behavior [125], whereas reflection on experience and self-efficacy were not found to be relevant. In a study of German secondary mathematics teachers at two measurement points, a causal effect of teacher self-efficacy expectations on subsequent instructional quality (self-reported teachers' self-efficacy and instructional quality) was partially found [126]. Given the primarily heterogeneous and partly contradictory findings, the need for research on self-efficacy or self-efficacy expectations in the context of diagnostic skills becomes clear.

1.5. Claim and Research Questions

In addition to content knowledge about the addressed context in an experiment and about experimentation (CK), the teacher needs pedagogical content knowledge about typical difficulties of students in the experimental implementation of the context as well as about possibilities for action in the experimental instructional setting (PCK) [127,128]. In this study, these dimensions of professional knowledge are addressed as important components in the formation of diagnostic competence. Consequently, the promotion of diagnostic competence requires that these areas of knowledge are either developed or are already present in the pre-service teachers. Especially in the first third of university studies, it is important to clarify what previous knowledge students bring with them or do not bring with them in the development of diagnostic competence for the assessment of students' skills to experiment or are not developed in the basic subject didactic training. On this basis, university teaching-learning programmes can be optimized and tailored to the target group. Since research studies in science education with a focus on the evaluation of diagnostic competence in combination with subject-specific pedagogical knowledge are rare so far [129,130], an explorative approach was chosen for the present study. In a chemistry course and a biology course, respectively, at two institutions at the University of Braunschweig and the University of Kassel the first step was to qualitatively investigate the professional knowledge of methodological difficulties and central decisions/actions of students in science experimentation among pre-service teachers in the first third of their university teaching studies. For this purpose, the following explorative-qualitative research questions were considered:

RQ1: Which difficulties in experimentation pre-service teachers with the subject biology and/or chemistry are able to describe on the basis of their pedagogical content knowledge (PCK) in the first third of their university studies?

RQ2: To what extent can pre-service teachers with the subject biology and/or chemistry describe central methodological contents and actions for experimentation in the first third of their university studies and what (methodological) content knowledge is predominant here (CK)?

A number of individual and contextual factors may influence the willingness and ability of (pre-service) teachers to implement diagnostic activities (while experimenting) in the teaching profession (cf., e.g., [125,131]). Attitudes toward the relevance and importance of diagnostic content in teacher education and, more broadly, the teaching profession are difficult to predict due to overlaps in the addressed knowledge domains. As a content element of pedagogical/didactic and/or educational study elements, both higher and lower attitude expressions would be expected according to the findings of Cramer [132]. Similarly, for the area of self-efficacy expectations with the specification of diagnostic competence in subject-related settings of experimentation, there is a lack of empirical

findings that can be used to make an educated guess about the expression in the sample studied here. Consequently, the following additional descriptive-quantitative questions were examined in order to draw statements about the expression of diagnosis-related attitudes and self-efficacy expectations among pre-service teachers:

RQ3: What attitudes towards the relevance of diagnostics in teacher training and what self-efficacy expectations for diagnostics in experimental settings show pre-service teachers in the first third of their university studies and how are they related in terms of expression?

RQ4: Is there a relation between pre-service teachers' attitudes towards the relevance of diagnostics in teacher training resp. self-efficacy expectations for diagnostic activities in experimental settings:

- a. . . . and their pedagogical content knowledge about difficulties of students in experimentation?
- b. . . . and their (methodological) content knowledge about central components and decisions in scientific experimentation.

2. Materials and Methods

2.1. Procedure

The written survey on which this paper is based was conducted at the beginning of two regular obligatory courses identified in the module plan from the 2nd semester of the (bachelor/teacher) degree programme in chemistry at the University of Braunschweig and biology at the University of Kassel. The procedure and information provided, as well as the time frame for completion (approximately 30 min) were identical in the cohorts. An online questionnaire was used to collect (a) demographic and academic information, (b) pedagogical content-related diagnostic knowledge about student difficulties and central actions or decisions of students in experimentation, and (c) the relevance of diagnostics in teacher training as well as domain-specific self-efficacy expectations. When recording the self-efficacy expectations, the participants were free to specify them in relation to the subject biology or chemistry. They had to make a selection beforehand and were assigned to the chemistry or biology sub-sample according to this selection (see Section 2.2).

Participation in the survey was anonymous and voluntary. For the release of the socio-demographic, quantitative, and qualitative data of the closed-ended and open-ended questions, the participants provided a declaration of consent for anonymous analysis and publication.

2.2. Participants

The sample consisted of 51 pre-service teachers of biology and/or chemistry. Of these, 66.7% were female and the average age was 22 ± 2.6 years. Of the participants, 34 were studying to become teachers at grammar schools and 17 were studying to become teachers at secondary schools. The relevant sociodemographic data for both the total sample and the sub-samples are presented in Table 2. In relation to the sub-sample, the proportions of pre-service teachers in their second or fourth semester were 47% in biology and 62% in chemistry. While a further 43% of pre-service teachers in biology were in their 6th semester, the remaining proportion in chemistry was distributed among the higher semesters. The average number of semesters is similar in both sub-samples, as is the distribution of gender and the type of school targeted in the study programme (see Table 2).

Table 2. Demographic characteristics of the participants.

	<i>n</i>	Age in Years	Sex		School Type		Semester in Biology	Semester in Chemistry
		<i>M</i> (<i>SD</i>)	Female <i>n</i> (%)	Male	GYM <i>n</i> (%)	HR	<i>M</i> (<i>SD</i>)	<i>M</i> (<i>SD</i>)
Total sample	51	22.0 (2.6)	34 (67)	17 (33)	34 (67)	17 (33)		
Sub-sample biology	30	21.9 (2.5)	22 (73)	8 (27)	19 (63)	11 (37)	4.7 (1.8)	
Sub-sample chemistry	21	22.2 (2.8)	12 (57)	9 (43)	15 (71)	6 (29)		4.4 (2.1)

Annotation: *n* = number; *M* = mean; *SD* = standard deviation; GYM = “Gymnasium” (grammar schools); HR = “Haupt-/Realschule” (secondary schools).

Approximately three-quarters of the participants had taken one or two courses in subject didactics at the time of the survey (biology 67%, chemistry 77%); all others had already taken three or more courses in subject didactics. At the University Braunschweig, the largest proportion of pre-service teachers recruited were those from the chemistry sub-sample. Only two participants of this sub-sample completed the survey in the context of a chemistry didactic course at the University Kassel. Further participants from this course could not be included in the analysis due to too high a semester number and missing data. Accordingly, a description of the university teaching and pre-conditions for the chemistry sub-sample focuses on the curricular structures at the University Braunschweig. Here, pre-service teachers take subject courses in the first semesters, in which investigations tend to follow detailed instructions in laboratory practicals. Content knowledge and skills in natural sciences working methods and techniques and handling laboratory materials are to be developed. In the 4th semester, pre-service teachers usually attend courses with didactic content for the first time, in which, among other things, the hypothetical-deductive processes for scientific inquiry and experimental problem solving are addressed. The data collection took place at the beginning of the seminar “Simple scientific experiments”.

The biology sub-sample mainly consists of pre-service teachers with biology as a subject from the University of Kassel. Here, too, there is an exception to two participants who come from the University of Braunschweig and who chose biology rather than chemistry in the survey section on domain-specific self-efficacy expectations. However, since more than 90% participated in the survey at the beginning of a course in the didactics of biology at the University Kassel, the curricular structures available here are used to describe the university teaching- and pre-conditions for the biology sub-sample. The first two semesters of the biology teaching programme at the University of Kassel are also dominated by subject-related biology courses (incl. laboratory). In addition, the basic module “Introduction to Biology Didactics” with lecture and exercise should be completed in this stage. In this module, two individual sessions provide basic background and information on the scientific inquiry process and the associated subject-methodological (procedural) knowledge. In each subsequent semester there are further subject didactic courses with different emphases. The survey took place at the beginning of the course “Scientific inquiry methods and lab techniques in biology teaching”, which is central to the content area of scientific inquiry; it can therefore not be assumed that the participants have received in-depth training in this area.

2.3. Instruments

Some authors critique a lack of validity evidence for instruments to assess scientific reasoning competencies (e.g., [133,134]) and point out that multiple choice assessment can hardly be seen as situations closely representing real life (e.g., [135]). In some studies with students response processes have also been examined qualitatively (through thinking aloud, eye-tracking studies, video recordings and written recordings), and these studies confirmed that respondents use procedural and epistemic knowledge (e.g., [136,137]).

Moreover, as science is constituted among others by specialized language [138], a central part of this study is a qualitative instrument to determine pre-service teachers' scientific reasoning knowledge concerning student' difficulties.

2.3.1. Student Difficulties/Misconceptions and Actions in Experimentation

The central concern of the open-ended questions posed here in two parts is to record the pre-service teachers' Pedagogical (didactic) Content Knowledge (PCK) about typical students' difficulties/misconceptions as well as their Content (methodological) Knowledge (CK) about the processing of an experimental task for an authentic teaching scenario. The latter is presented to the subjects in the form of a short progression plan from the perspective of the teacher and the experimental task given to the students via a task sheet (see Supplementary Materials S1). The phenomenon to be investigated here for dissolving sugar is visualized to the (fictitious) students as well as to the pre-service teachers by a video with a dialogue between two friends at and about a cup of tea ('tea conversation') and used to derive the question "How does the time needed to dissolve sugar depend on different influencing factors?" This task stem is followed by two open-ended questions or subtasks. In part 1, the central difficulties of the students in accomplishing the set experimental task or the planning and execution of an experiment on the dissolving time of sugar are examined with regard to the content knowledge and the experimental implementation. Part 2 explicitly aims to attain methodological (procedural) knowledge for the design of experiments in the sense of the scientific inquiry process and to also indirectly overcome possible obstacles to master this process for students. In this part, four central decisions or necessary actions of the students to accomplish the experimental task are to be articulated by the pre-service teachers in writing.

2.3.2. Relevance of Diagnostics in the Teacher Training Program

The measurement instrument for assessing attitudes toward the importance of diagnostics in teacher education includes five items, the language and content of which were adapted to fit the focus of the present study, based on Lorenz [139]. Based on the characteristic values obtained in this sample for Cronbach's alpha [140] (pp. 281–302) with 5 items, $\alpha = 0.64$, and a low discriminatory power for the item: it is important for teachers to be able to correctly assess students' performance in experimentation ($r_{it} = 0.165$), a reduction of the scale to four items is made ($\alpha = 0.70$). A complete overview of all items can be found in Appendix A.

2.3.3. Domain-Specific Self-Efficacy Expectations

Based on the concept of self-efficacy as devised by Schwarzer and Jerusalem [141], an instrument was developed for the domain-specific assessment of subjective certainty in coping with diagnosis-related teaching activities in general, as well as in instructional experimental settings. In addition to an explicit focus on this selected activity domain, the item formulation describes the perceived ability to be assessed here as "proficiency". Moreover, the ability of an action is specified in its effectiveness via the inclusion of challenges or obstacles or barriers to action in the item formulation [109]. The items used in the present survey [142] were administered and piloted to 98 pre-service teachers majoring in biology over three semester cohorts. Based on the piloting data, the dimensionality of the scale was tested using exploratory factor analysis (principal component analysis with varimax rotation) and internal consistency was tested using Cronbach's alpha. Factor analysis revealed two factors with an eigenvalue > 1 that explained 59% of the variance. On the first factor, four items loaded highly (factor load (a_{jq}) > 0.649). One additional item loaded similarly on both factors; based on content, this item was assigned to the second factor. Thus, factor one (experiment-related diagnostic activities) is represented by four items whose content is explicitly diagnostic and/or experimental. On the second factor, another three items loaded similarly highly ($a_{jq} > 0.709$). Together with the content-related item, this factor describes diagnostic activities implicitly and without the inclusion of

experimentation/processes (general diagnostic activities). The reliability of the two scales formed according to the two-factor model was in the good range ($\alpha \geq 0.70$). The items and (factor analytic) findings for piloting the instrument on domain-specific self-efficacy expectations can be found in Appendix B.

In the survey presented in this article, a further subject-related specification took place in the item formulation and content orientation. In accordance with the sub-samples, the subjects were provided with either biology or chemistry items to assess their self-efficacy expectations. Using a 4-point Likert scale ranging from “does not apply at all” (1) to “fully applies” (4), the pre-service teachers’ self-reports were recorded in relation to the respective items. The empirically derived and validated scales from the pilot were adopted; the biology sub-sample showed in the very good range ($\alpha \geq 0.83$) and the chemistry sub-sample showed acceptable Cronbach’s alpha values ($\alpha \geq 0.61$).

2.4. Data Analysis

2.4.1. Qualitative Analysis Methods

The written student responses on knowledge about student difficulties and actions/procedures during experimentation were analysed qualitatively according to the procedure of a summarising content analysis with deductive-inductive category formation [143] using the programmes MAXQDA 2020 and Excel 2016. For the first part with written pre-service responses on student difficulties and misconceptions, the formation of content categories was initially carried out deductively with a theoretical foundation based on research findings, which primarily explicate the prior knowledge of the researchers involved in the study [39,47], as well as on results from a wide range of literature sources (see Table 1). Where possible, the references were classified in the process-related sub-steps of experimentation in Mayer’s [34] structural model of scientific reasoning, whereby the formulation of the question was already specified in the material. In the second part (procedure for experimenting), the central decisions listed were sorted into subcategories inductively formed on the material on the basis of similarities in content. These were then classified by subsumption into supercategories based on the hypothetical-deductive procedure in the experiment [34]. This systematic inclusion of deductively and inductively formed categories served to identify and explore further meaning components in the process of creating the category system up to the category definition through anchor examples [144]. The assignment of the student responses to the categories for the 1st and 2nd part was carried out independently by two trained raters. Interrater reliability was estimated using Cohen’s Kappa with the programme IBM SPSS Statistics (version 27) and is ‘substantial’ in the 1st part (Cohen’s $\kappa = 0.80$; $p \leq 0.001$) and ‘almost perfect’ in the 2nd part (Cohen’s $\kappa = 0.903$; $p \leq 0.001$) [145]. Finally, the relative frequency for each code in each category was calculated for each group of biology or chemistry students.

2.4.2. Quantitative Methods of Analysis

In order to analyse the attitudes towards the importance of diagnostics in teacher education and domain-specific self-efficacy expectations towards diagnostic skills/actions in teaching-learning settings for science experimentation, descriptive procedures, representations, and associated characteristic values were used and analysed with the programme IBM SPSS Statistics (version 27). In the first step, deductively derived as well as newly constructed items for the assessment of self-efficacy expectations were used in a pilot sample in order to test them accordingly factor-analytically and in their reliability (see Section 2.3.3). A further reliability test was also carried out for the two empirically derived scales on domain-specific self-efficacy expectations as well as for the newly constructed scale on the significance of diagnostics in teacher training in the sample on which this study is based. Taber [146] was used to assess the Cronbach’s for internal consistency. In the next step, the sample was analysed in the subjects included here and the constructs examined in each case (Mann–Whitney U test; Wilcoxon test) in order to uncover any differences in attitudes and self-assessments that could influence possible correlations

with the results on the open-ended task and allow conclusions to be drawn about subject specifics. According to Cohan [147], possible effects associated with this are rated as insignificant for $r < 0.10$, weak for $r = 0.10\text{--}0.30$; medium for $r = 0.30\text{--}0.50$ and strong for $r > 0.50$. In accordance with the goal of describing a comprehensive picture of the extent of subject-specific methodological (procedural) knowledge of experimentation in combination with diagnosis-related competence assessments in this area, frequencies and mean values in the respective characteristics are reported and, in the last step, the correlations between the quantitative and qualitative data are exploratively tested via correlations (Spearman rank correlation). Due to the small samples that predominate, especially in the subject sub-samples, and a violation of the criterion for normal distribution in selected scales, non-parametric procedures were used throughout.

3. Results

In total, 49 fully completed questionnaires by the pre-service teachers for the first sub-task (student's difficulties) and 50 for the second subtask (procedure for experimenting/key decisions) were included in the predictive diagnostics.

(RQ1) To analyze participants' responses to the first subtask, process-related difficulties and misconceptions among students described in the theoretical literature (see Table 1) were applied to the material. The pre-service teachers' responses were paraphrased prior to analysis, which involved rewriting the responses' core content in a concise descriptive form [143]. This then allowed them to be reliably assigned to the categories. For example, the statement "... The students do not know what things are important/what one needs to pay attention to—no connection to the research question—no hypotheses, ..." was reformulated into the core components "creating a link to the research question" and "no hypothesis generated". The statement "... should be clarified and potentially what a conjecture/thesis entails" was reduced to "technical term conjecture/thesis not known". A total of 293 statements could be coded from the 49 answer sheets examined. Of these, 166 statements referred to process-related difficulties and/or misconceptions by students, which corresponds to roughly 57% (Table 3). In addition, the respondents frequently mentioned difficulties arising from the "instructional setting", specifically from the open-ended task structure and materials pool, which overwhelmed students and required them to make decisions. Statements referring to students' decisions about how to divide up tasks and/or disagreements within the group were assigned to the category "social format". Statements referring to teachers' instructional planning were assigned to the category "teacher". A total of 94 statements (32%) were assigned to these non-process-related categories. Difficulties related to subject-specific content knowledge or technical terms were mentioned in 33 statements (11%).

In the following, the difficulties that students have to deal with during each sub-steps within the hypothetico-deductive scientific inquiry process according to the pre-service teachers' statements are explained (see Table 3).

In terms of the **phenomenon**, the pre-service teachers mentioned a lack of or differences in prior knowledge. (The students will not yet be familiar with the phenomenon of diffusion; [...] was not worked through as a group, meaning that the students will not be able to think deeply about the phenomenon. It also seems that no ideas were taken up or discussed. Hence, there is no bridge to their prior knowledge [...]).

Potential difficulties in dealing with the provided **research question** to be investigated in the fictitious experimental instruction setting (see Supplementary Materials S1) were also identified. These difficulties concerned the students failing to understand or refer back to the research question ("The students might not be able to make connections between the conducted experiment and the research question").

Table 3. Overview of the students' process-related misconceptions and difficulties faced during experimentation identified by the pre-service teachers (incl. paraphrases) according to the three phases for scientific reasoning of the SDDS model [22].

Difficulties Students Face (Deductive, See Table 1)	Paraphrase from Material	<i>h</i>
Phase: Search Hypothesis		
Phenomenon		
Different levels of prior knowledge	Students have sharply divergent understandings (or sometimes no understanding) of how sugar dissolves in tea.	13
Research question		
Integrating the research question into the experimentation process	Perhaps a too complicated research question that might lead to misunderstandings during implementation. As a result, there is no bridge to students' prior knowledge and the students cannot incorporate their contributions into the research question.	8
Hypothesis Generation		
No hypothesis generated	No hypotheses.	1
No understanding of hypothesis generation	No background subject-related knowledge is present, nor is fundamental knowledge of scientific knowledge acquisition through the generation of hypotheses.	2
Suppositions not linked to research question	[...] generate a hypothesis that makes reference to the research question and subsequently guides the experimentation phase.	1
Phase: Planning and Testing		
Planning		
Selection of materials unsystematic, incomplete and/or via trial-and-error	[...] that they throw together materials at random. [...] there are still problems with the selection of materials, since typically only the necessary materials are made available and the students try to use everything, even when it's not necessary.	19
Trying things out in an unstructured way (no plan, change all variables)	Confusion with respect to materials selection, since more materials are available -> perhaps the students want to switch to using other materials during the experiment. [...] change their minds while conducting the experiment if they get the feeling they have selected the "wrong" factor.	3
Planning an experiment that (does not) achieve its objective	No foundation for planning an experiment.	3
Dealing with variables in an unsystematic way (multiple variables are confounded or a single variable is varied unsystematically)	Students conduct the experiment with two independent variables simultaneously, which does not lead to an unambiguous result.	19
Aware of confounders, but do not eliminate them	When conducting the experiment, it could be difficult to stick to one influencing factor and not hold constant conditions constant. That they do not recognize or eliminate confounders when constructing a self-developed experiment.	2
Lack of control group approach	Furthermore, students often forget to create blinded or comparison samples.	2
Variable selection		
Deciding which variables to include	[...] which influencing factors play a role in the solubility of sugar in water. [...] which experiment should investigate a certain influencing factor.	42
Conducting the experiment		
Variables operationalized not at all or incorrectly	Imprecision in keeping time, since the time point at which all of the sugar has dissolved is often difficult to determine.	14
Imprecise measurement/Measurement error	[...] that they measure the water by eyeballing it or count without looking at a clock.	
Order of experimental steps	[...] students probably do not yet know exactly how an experiment with a phenomenon, planning, observation and interpretation is structured. I consider difficulties in conducting the experiment graver here [...].	12
Conducting experiment incorrectly	Experiment not conducted in a structured way.	6
Documentation	Measurements not recorded.	3

Table 3. Cont.

Difficulties Students Face (Deductive, See Table 1)	Paraphrase from Material	<i>h</i>
Phase: Evaluate Evidence		
Unexpected data are attributed to errors in conducting the experiment	They might change their results when they have the feeling that something is not right or they have done something incorrectly.	1
Insufficient reflection on the experimental results	Errors are not taken into consideration.	1
Replicability/validity	[...] replicability or precision of measurement will pose problems for the students.	2
Wrong conclusion drawn from consistent experiments	The students might not be able to correctly interpret their observations.	1
Conclusions are not possible because variables were not varied systematically, and/or fallacious conclusions are drawn	[...] not investigate multiple influencing factors at the same time, since they then cannot say which factor actually influences the sugar's dissolution. Analysis does not take place, causal conclusions are not possible.	10
Data are not related (back) to hypothesis/research question	[...] hypotheses are actually confirmed or falsified based on the insights gained while conducting the experiment.	1

Annotation: *h* = absolute frequency of mentions from a total of *n* = 49 pre-service science teachers.

Only a few statements mentioned difficulties regarding **hypothesis generation** compared to the later steps of the process. Three participant statements referred to the aspects of having no hypothesis, hypothesis formulation, and generating multiple hypotheses, while one statement referred to the link between the supposition and research question. No pre-service teachers mentioned formulating justifications for hypotheses as a difficulty faced by students. One pre-service teacher wrote: "To me, the connection between technical terms and phenomena observed in everyday life does not seem to be pronounced enough in sixth grade in order to bridge the gap from multiple hypotheses [...] to independently conducting an experiment to test these hypotheses". This statement addresses the association between the hypothesis and the need to plan an experiment that successfully tests the hypothesis.

With regard to **planning**, no respondents mentioned potential problems connecting the hypothesis and experiment, i.e., planning an experiment that is able to actually test the proposed hypothesis. In contrast, difficulties with planning a meaningful experiment and selecting appropriate materials (e.g., utilization of the materials pool) were mentioned very frequently ("It is possible that the students might become overwhelmed by the various materials on offer"). Application of the control-of-variables strategy was also described by many respondents ("Moreover, various factors might be unwittingly coupled with one another, meaning that only a single factor is not investigated"). Three mentioned difficulties concerned trying things out in an unstructured way ("no plan-change all variables") ("that the students change their minds while conducting the experiment if they get the feeling they have selected the 'wrong' factor").

Variable selection was included as a unique overarching category, since it accounted for 25% of process-related statements and was mentioned by 28 participants ("It could be the case that no influencing factors are selected, but only changes that exert no influence"; Ultimately, the students need to know and/or be able to identify the influencing factors in order to independently develop an experimental approach"; "A difficulty for students would be recognizing the difference between loose sugar and sugar cubes").

With respect to **conducting the experiment**, general problems were mentioned: "Difficulties arise in conducting the experiment". The spectrum of statements referring to students' difficulties with respect to failing to or incorrectly operationalizing variables, from imprecise measurements to measurement errors, ranged from taking imprecise measurements with the stopwatch to the use of different amounts and/or volumes of the

chemicals and measurement differences between students (“When conducting the experiment, replicability or precision of measurements will pose problems for the students.”; “constantly switch the person doing the measurement, which can lead to imprecise measurements”). Students’ difficulties regarding documentation also fall under this category (“that the results need to be documented; the learners might not do this and forget the times they measured, for example”). With respect to evaluating evidence, only one respondent mentioned that students might attribute unexpected results to an error made in conducting the experiment (“They might change their results when they have the feeling that something is not right or they have done something incorrectly.”). Insufficient reflection by the students on the experimental results was mentioned once (“Errors are not considered—No analysis takes place (causal conclusions are not possible).”). The most frequently mentioned difficulty was that no conclusions are possible because the variables were varied unsystematically (“Students conduct the experiment with two influencing factors simultaneously, which does not lead to an unambiguous result”; “Focusing on just one aspect is probably difficult for some groups of students, so they try to investigate multiple factors simultaneously and only realize at the end that they cannot draw any conclusions from the investigation”).

Problems with **evaluating evidence** due to lack of a comparison or blinded sample were also mentioned (“Furthermore, students often forget to create blinded or comparison samples and thus cannot draw any concrete conclusions from their results”).

Overall, these problems were attributed to students being overwhelmed by the “lack of guidance” in the self-directed procedure, which did not involve following prescriptive, externally prescribed experimental steps (“Moreover, it is not precisely described how the experiment should proceed”; “that the students can become overwhelmed by this freedom and autonomy”). These statements were assigned to a non-process-related overarching category, the “instructional setting”, which the pre-service teachers referred to 71 times, with particular focus on obstacles stemming from the open-ended nature of the task. Additionally included in this category was students’ lack of experience in dealing with experimental materials, which was mentioned 10 times ([...] difficulty correctly using the given materials”; [...] that the students are not familiar with all the materials and how they are used”). Overall, 94 statements were coded into one of three such non-process-related categories, the “instructional setting”, “social format” and “teacher”. Twelve statements concerning students’ decisions about how to divide up tasks and/or disagreements within the group were assigned to the category “social format” (“Students cannot come to an agreement within the group”). Eleven statements addressing the teacher’s instructional planning were assigned to the “teacher” category (“[...] handouts to assist students or opportunities to repeat explanations are not included in the experiment, which could lead to excessive questions”; “Since no thermometer is available, the students cannot make any statements about the water temperature”). Beyond these overarching categories and subcategories, difficulties related to subject-specific content knowledge and technical terms were also reported. Lack of familiarity with technical terms was mentioned 14 times, with the term “influencing factor” coming up seven times, “conjecture” three times and other technical terms four times (“have problems finding out what influencing factors are exactly”; “even the term ‘influencing factor’ should be clarified and possibly also what a conjecture/thesis entails”) (“Effect on dissolution speed is difficult in sixth grade”). That subject-specific content knowledge might be lacking, insufficient or fragmentary was mentioned 19 times (“With respect to subject-specific content knowledge, the difficulty might arise that some students do not know that sugar dissolves more quickly and easily at higher temperatures”; “Background subject-related content knowledge is not yet present”).

To summarize, the respondents considered here ($n = 49$) mentioned four of the nine overarching categories on average ($M = 4.00$, $SD = 1.26$; $\min = 2$; $\max = 7$) in their written responses asking about potential difficulties faced by students, including roughly two of the five process-related overarching categories (phenomenon, hypothesis generation/research question, planning, variable selection, conducting the experiment, and evaluating evidence)

($M = 2.33$, $SD = 1.18$; $\min = 0$; $\max = 4$). Three pre-service teachers did not mention any process-related difficulties. Statements falling under the overarching category of hypothesis generation were by far the least frequent (2% of the total number of statements in the process-related categories). Considered together with the phenomenon and the research question, this rose to 10% of all statements in the process-related categories, which is similar to the share of statements referring to evaluating evidence (15%). Planning was mentioned most frequently by the respondents (29% of all statements in the process-related categories), with the lion's share referring to selecting appropriate materials and the control-of-variables strategy, each of which made up 11% of all statements in the process-related categories.

Differences in the quality of participants' statements were related to their subjects of study, with the pre-service chemistry teachers using technical terms like "familiarity with the RGT rule" (authors' note: Reaction velocity-Temperature-Regulation), "materials surface", "solubility product", and "law of mass action" more frequently than the pre-service biology teachers, who used more general formulations like "dissolution time depends on the temperature", "recognizing the difference between loose sugar and a sugar cube", "factors like the solubility or saturation of a liquid and corresponding effect on the dissolution speed". The only further differences between the biology ($Mdn = 6.00$) and chemistry pre-service teachers ($Mdn = 4.50$) uncovered concerned the total number of difficulties identified in all nine overarching categories (Mann-Whitney U-test: $z = -2.377$, $p = 0.017$; $r = 0.34$). Examining the frequencies of the process-related and non-process-related categories revealed that this difference reflected a larger share of non-process-related categories in the biology pre-service teachers' statements. Consequently, no significant differences in the number of process-related difficulties mentioned were found. A total of 24% of the biology pre-service teachers ($n = 29$) referred to the groups' social fabric as it related to completing the tasks, compared to just 10% of the chemistry pre-service teachers ($n = 20$). A similar pattern was found for the "teacher" category, which was mentioned by 24% of biology pre-service teachers but only 5% of chemistry pre-service teachers.

(RQ2) The university pre-service teacher statements about four key decisions students need to make when solving problems experimentally could be assigned to seven superordinate categories and 26 subcategories (Table 4). A total of 325 statements were analyzed, of which 298 were assigned to the following process-related superordinate categories, which referred to phases of the experimentation process [22,34]: decisions about the "phenomenon", the "research question and/or hypothesis", "working with and identifying variables", "planning", "conducting inclusive documentation", "analysis and interpretation". All statements by participants were analysed, regardless of the number of decisions a respondent mentioned, which ranged from 2 to 14 in total (both process-related and non-process-related).

On average, the pre-service teachers mentioned decisions about a bit more than three (3.34) process phases/elements (=superordinate categories). In addition, five pre-service teachers made six mentions of decisions by teachers, such as ensuring appropriate group composition or an appropriate task. None of the participants mentioned all of the process phases/elements considered here in their responses. Turning to the subcategories, the pre-service teachers most frequently mentioned decisions related to the influencing factors (around 20% of the 298 total mentions of process-related decisions) and the experimental materials (12%), in line with the difficulties students were considered to face in this area (see Table 3). Among decisions about "variables" and "planning", the aspects of avoiding confounders, mentioned twice by two different pre-service teachers, and including a control group, mentioned three times by two different pre-service teachers, were grossly underrepresented. This was also the case for estimating the experimental validity within the superordinate category of "analysis and interpretation", which was mentioned by three pre-service teachers one time each.

Table 4. Overview of categories (superordinate and subcategories) in the statements about key decisions students face when solving problems experimentally, as well as corresponding anchoring examples and number of mentions.

Superordinate Category Decisions About	Subcategory	Anchoring Example	<i>h</i>	<i>f</i>
... the phenomenon	Retrieving (prior) knowledge about the phenomenon	The students need to visualize the phenomenon and think about how they can link it to their prior knowledge.	4	
			4	8%
... the research question and/or hypotheses	Develop a research question/hypothesis	To start with, the students should think about a research question (or a conjecture) that they can subsequently answer with the experiment. They need to come to agreement amongst themselves on how they can best express the characteristic to be observed and in what way they will investigate it.	17	
	Working with the research question/assigned task		2	
	Understanding the research question	Students need to understand the research question.	1	
			20	36%
... working with and identifying variables	Selecting one influencing factor	The students need to decide which influencing factor they want to test.	59	
	Controlling for other factors	Investigate one factor in the experiment! [...] pay attention to other factors that remain constant.	12	
	Decisions about the measured variable	It needs to be determined when the time will be stopped (when has the sugar dissolved?).	8	
	Avoiding confounders	[...] As part of this, the students must minimize the presence of confounding factors.	3	
	Decisions about the dependent variable (degree of breakdown, amount)	They need to agree on whether to use ground sugar or sugar cubes.	18	
			100	86%
... planning	Selection of appropriate materials	Decision about selecting materials from the list of materials.	36	
	Planning the experiment	[...] decide how they will proceed.	17	
	Determining the order of work steps	The learners need to familiarize themselves with which work steps they will conduct in which order.	4	
	Planning a control group experiment	At least one comparative experiment must be conducted.	4	
	Replication/Reliability	Conduct each experiment at least twice.	4	
			65	84%
... conducting the experiment (including documentation)	Conducting the experiment	Conduct the experiment in accordance with the experimental plan.	24	
	Documentation	Selecting documentation of the experiment.	21	
	Observation	The students now observe [...].	15	
			60	70%
... analysis and interpretation	Conclusion	They need to interpret their results (conclusion). Based on what evidence they can see how the influencing factor they are investigating affects the sugar's dissolution.	17	
	Evaluating evidence/causal relationships	[...] pay attention to errors that might have crept in and potentially conduct the experiment again.	8	
	Reflecting on errors	Test the hypothesis	7	
	Referring back to the hypothesis/research question	Analysis (Was the hypothesis refuted or not?)	6	
	Paying attention to validity	Does the experiment I have conducted actually answer my question?	9	
			2	
			49	50%

Table 4. Cont.

Superordinate Category Decisions About	Subcategory	Anchoring Example	<i>h</i>	<i>f</i>
Decisions unrelated to the scientific method	Dividing up tasks within the group	They need to divide up the various tasks within the experiment [. . .]	11	
	Completing the group work together	The group should work together so that everyone is aware of what has been done.	9	
	Working precisely and carefully	The students need to work very precisely.	3	
		Time management: "The time allotted should also be considered."	2	
	Other	Cleaning up after the experiment	1	
	Alignment with teacher (minimizing errors)	1		
			27	38%

Annotation: *h* = absolute frequency of mentions from a total of *n* = 50 pre-service science teachers; *f* = relative frequency of mentions in the subordinate category (%); each respondent could identify multiple influencing factors.

Overall, the pre-service teachers frequently adopted a results-focused perspective: "[...] After conducting the experiment, the students should collect the results, discuss them as a group [...]"; "How can I answer my question for others. How can I be sure of my result, does the experiment I have conducted really answer my question". Around half of the pre-service teachers adopted an understanding-oriented perspective and explicitly mentioned the need for students to "understand the research question/assigned task" and the need to make decisions about "selecting appropriate materials" or the "order of work steps". In this context, references were made to the open-ended structure of the task, the third most commonly mentioned difficulty for students in Part 1 (see Table 3), ("The learners need to familiarize themselves with which work steps they will conduct in which order"; "... it could be difficult if they do not have sufficient practice in conducting experiments and the students forget steps like documenting the results"; How do I conduct the experiment, which work steps logically follow one another") as well as decisions about selecting materials from the materials pool ("They need to ascertain what materials are necessary to match the selected influencing factor"; "They need to decide what materials they want to use and which ones they don't need or don't want to use"). Three pre-service teachers focused exclusively on deciding on an influencing factor, such as water temperature or water volume, or deciding whether to stir or shake the samples during the experiment.

When comparing the two subsamples, it was found that the biology pre-service teachers (Mdn = 4.00) made reference to more process steps (=superordinate categories) when discussing decisions that need to be made in carrying out the example experiment than the chemistry pre-service teachers (Mdn = 3.00; Mann-Whitney U-Test: $z = -2.639$, $p = 0.008$; $r = 0.37$). A similar pattern was found for the total number of process-related decisions across all subcategories among the biology (Mdn = 6.00) and chemistry pre-service teachers (Mdn = 5.00). The biology students reported more unique process-related decisions here as well; the difference just failed to reach significance and represented a small effect ($r = 0.27$; Mann-Whitney U-Test: $z = -1.931$, $p = 0.056$).

(RQ3) The university pre-service teachers' attitudes toward including diagnostic elements and promoting diagnostic skills in university teacher education were positive and quite pronounced. The full-sample mean was substantially above the scale midpoint, at $M = 3.13$ ($SD = 0.395$), (see Appendix B). No significant differences in the strength of these attitudes were found between the two subsamples ($Mdn_{\text{Chemistry}} = 3.00$; $Mdn_{\text{Biology}} = 3.25$; $z = -1.414$, $p = 0.157$). Likewise, there were no significant differences between the two subsamples of pre-service teachers in perceived domain-specific self-efficacy with respect to general ($Mdn_{\text{Chemistry}} = 2.50$; $Mdn_{\text{Biology}} = 2.50$) as well as experimentation-related diagnostic activities ($Mdn_{\text{Chemistry}} = 2.50$; $Mdn_{\text{Biology}} = 2.25$). However, there was a significant difference in the full sample between the two perceived self-efficacy scales and attitudes towards the importance of diagnostics in university teacher education. The

pre-service teachers' subjective perception of their own skills in successfully carrying out general and experimentation-related diagnostic activities ($Mdn_{\text{Self-efficacy factor1/2}} = 2.50$; $Mdn_{\text{Relative importance of diagnostics}} = 3.00$) was lower than their perception of the importance of this teaching and learning topic for university teacher education (Self-efficacy factor 1: $z = 5.612$, $p < 0.001$; Self-efficacy factor 2: $z = 5.345$, $p < 0.001$; $n = 50$). The effect sizes can be considered large, $r > 0.75$. However, no correlation between these two motivational constructs was found. The pre-service teachers saw potential for improvement in the diagnostic competences referred to in the perceived self-efficacy scale, as they tended to "somewhat disagree" or "somewhat agree" to these items (e.g., $M_{\text{Self-efficacy factor1-Bio}} = 2.38$, $SD = 0.61$; Table 5).

Table 5. Item scores for the domain-specific perceived self-efficacy scales.

Item	Biology ($n = 30$)		Chemistry ($n = 20$)	
	M (SD)	r_{it}	M (SD)	r_{it}
<i>Experimentation-related diagnostic activities</i>				
Even when I am experiencing stress, I can still diagnose students' errors in experimentation in biology/chemistry class well.	2.20 (0.71)	0.708	2.40 (0.60)	0.524
I am certain that I am able to recognize children's specific difficulties in experimentation in biology/chemistry even when under a large amount of time pressure.	2.40 (0.77)	0.737	2.75 (0.44)	0.447
In biology/chemistry, I am able to accurately assess my students' level of learning prerequisites, even when I have little time available.	2.43 (0.63)	0.603	2.65 (0.50)	0.380
In biology/chemistry, I am able to accurately assess my students' experimentation skills, even when I have little time available.	2.37 (0.72)	0.803	2.45 (0.51)	0.430
	$M_{\text{scale}} = 2.35$ $SD_{\text{scale}} = 0.60$ Cronbachs $\alpha = 0.86$		$M_{\text{scale}} = 2.56$ $SD_{\text{scale}} = 0.36$ Cronbachs $\alpha = 0.66$	
<i>General diagnostic activities</i>				
I am able to successfully integrate diagnostic activities to accompany learning in my biology/chemistry instruction, even when I am under time pressure.	2.13 (0.78)	0.628	2.20 (0.62)	0.217
Despite a high level of heterogeneity, I am able to create tasks in biology/chemistry that allow me to appropriately check both weaker and stronger students' knowledge levels.	2.57 (0.77)	0.665	2.50 (0.69)	0.546
I am able to successfully take into account students' learning processes when formulating individual learning goals in biology/chemistry, even when these differ markedly.	2.47 (0.78)	0.743	2.35 (0.67)	0.440
In biology/chemistry, I am able to accurately assess my students' thought and work processes, even when I have little time available.	2.37 (0.67)	0.576	2.55 (0.51)	0.393
	$M_{\text{scale}} = 2.38$ $SD_{\text{scale}} = 0.61$ Cronbachs $\alpha = 0.83$		$M_{\text{scale}} = 2.40$ $SD_{\text{scale}} = 0.42$ Cronbachs $\alpha = 0.61$	

Annotation: n = sample size, M = mean, SD = standard deviation, r_{it} = discriminatory power.

(RQ4) Quantifying the qualitative findings into frequency scores for the difficulties students face and the different process-related decisions/actions during experimentation made it possible to test the associations between these and self-efficacy expectations and attitudes. (a) With respect to the first (qualitative) subtask, the ^anumber of difficulties mentioned in all superordinate categories, number of difficulties in the process-related superordinate categories, and number of process-related superordinate categories addressed were included in the analysis. There are no significant correlations between these knowledge-based expressions and attitudes toward the importance of diagnostics (e.g., ^aSpearman's $\rho = 0.005$, $p = 0.974$). The same is found in comparison to the domain-specific self-efficacy expectations. In this sample, the expression of self-efficacy expectations with respect to general as well as experimentation-related diagnostic activities is not related to the pre-service teachers' knowledge of student difficulties in experimentation. (b) For

the second (qualitative) subtask, the ^bnumber of process-related superordinate categories addressed in the mentioned decisions and the number of decisions mentioned in the process-related superordinate categories were included. These results also did not correlate with perceived domain-specific self-efficacy or ^battitudes towards the importance of diagnostics in university teacher education in the present sample (e.g., ^bSpearman's $\rho = 0.038$, $p = 0.790$).

4. Discussion

What began in the 1990s with a call for “Science for All” [148] has led to the setting of obligatory educational goals in countries’ and states’ curricula and standards regarding scientific inquiry and scientific reasoning as a component of scientific literacy (e.g., [4–6,149]). In this respect, school curricula in the natural sciences (e.g., NGSS [149], University location federal state 1 [150,151], University location federal state 2 [152,153]) require students to be able to conduct and reflect on scientific inquiry processes. Subject-specific content on promoting scientific reasoning in the classroom are also anchored in curricular standards for teacher education (e.g., [9,11] and thus must be taught in university teacher education. Appropriate teaching-learning concepts for scientific inquiry in university education can give future science teachers a better understanding of the difficulties and the (mis)understandings, alternative ideas or misconceptions students experience during experimentation, in order to be able to include those diagnosis-related aspects. Accordingly, there are lines of research focusing on (pre-service) teachers, e.g., measuring and assessing pre-service teachers scientific reasoning competencies in higher education [133,154,155], verification of validity [156], evaluation of translated versions [157], but these are rather limited and predominantly of a quantitative nature (e.g., [158]). Many studies also focus predominantly on developing and testing concepts and materials to promote subject-specific, pedagogical content knowledge and pedagogical knowledge related to scientific reasoning (e.g., video vignettes as a support for scientific reasoning, including video vignettes as a tool to promote students’ learning: e.g., [48,98,99,159]; seminar concepts: e.g., [160]). However, the first step is to identify pre-service teachers’ knowledge state in order to appropriately adapt university courses and to develop and/or employ alternative teaching approaches. Consequently, the present study did not focus on developing teaching-learning concepts for the subject didactic training of pre-service teachers; instead, the primary focus lies on potential prerequisites for learning such diagnostic activities, not only in the knowledge areas mentioned (CK; PCK), but also in terms of attitudes and self-efficacy regarding diagnostics and scientific inquiry (with a focus on experimentation). To summarize, it became clear that the pre-service teachers in the present study were only able to identify and cite a varying yet small number of potential student difficulties in the experimental problem-solving process from the comprehensive catalogue, presumably based on their own experience in school and basic training in subject didactics. The same was true of procedural knowledge in carrying out an investigation of an illustrative experimental phenomenon (“knowing how” [29,30]). In contrast, the teacher education students had a strong sense of the importance of diagnosing students’ experimentation skills and the hurdles students face, as well as a moderate level of self-efficacy in carrying out such diagnostic activities. The research questions underlying these findings can be answered and discussed in detail as follows:

(RQ1) The pre-service teachers cited difficulties students face in all steps of the experimentation process (see Table 1), but with very different frequencies (see Table 3). They predominantly described difficulties in the planning phase, including variable selection, followed by the implementation phase. Contrary to the usually higher self-assessment of scientific reasoning abilities [133], only a small number of unique difficulties were described with respect to evaluating evidence and most of all, formulating a research question and hypothesis, and the number of participants mentioning these areas was very low as well. With respect to the research question, this is possibly due to the structure of the task and teaching scenario applied in the study, in which students were provided with an

overarching research question (see Supplementary Material S1). Moreover, it can be concluded that pre-service teachers not only lack pedagogical (didactics) content knowledge about potential student difficulties in these phases, but also exhibit a lower level of content methodological (procedural) knowledge, e.g., with respect to formulating hypotheses. Students in the early semesters of university teacher education are still largely unfamiliar with the procedure and content of hypothesis formulation [155], which is at least partially due to the lack of or minimal opportunities to learn how to formulate research questions and hypotheses in school-based science instruction [42].

Overall, the pre-service teachers in the present study frequently attribute students' difficulties to the instructional setting, e.g., the open-ended nature of the task, working with the experimental materials, coming to an agreement and dividing up roles within the group during experimentation are most frequently mentioned. The participants' knowledge of PCK with regard to potential student difficulties and misconceptions hardly extends to aspects of subject-specific methodological concepts, such as planning an experiment that actually tests the hypothesis(es), the question of confounding variables or dealing with unexpected data. Overall, the respondents used rather general terminology, with scientific terminology [161] used only to a very small extent. Comparing the sub-samples, however, the pre-service chemistry teachers used technical terms more frequently, which may be due to the specific experiment selected in the example teaching unit (sugar's dissolution time) at the university location Braunschweig at the time of the survey. In contrast, it can be seen that the pre-service biology teachers cite more difficulties students face than the pre-service chemistry teachers. However, this difference is only due to a higher proportion of non-process-related categories mentioned by the pre-service biology teachers. As with the use of scientific terminology, this may be due to the number of subject didactics courses already completed by each group at the universities in which the study was conducted. Even though the study was conducted before any courses on scientific reasoning at both locations, the pre-service biology students in Kassel had already briefly dealt with this content area in their introductory lecture and tutorial.

(RQ2) With regard to key decisions (procedure for experimenting), the university pre-service science teachers tended to return to aspects they had also mentioned with respect to difficulties and misconceptions. The majority of respondents referred to decisions about working with and identifying variables in the experiment, planning, and actually conducting the experiment (including documentation). Interestingly, however, more than a third of the pre-service teachers also mentioned decisions about the research question or hypothesis, i.e., experimental steps that were underrepresented in the responses to the first subtask. It is possible that the pre-service teachers made greater reference here to hurdles had experienced themselves in their practical laboratory training at university. For example, one participant writes: "Now the students should not be thrown into the experiment like that. Forming hypotheses is often still a sticking point in tasks, even for university students, albeit at a higher level. [...] However, gaining knowledge should always be the goal of the experimental phase. Finally, in order to consolidate knowledge, the hypotheses must be verified or falsified based on the knowledge gained during the experiment." Results- and understanding-oriented statements are also made here, presumably due to the fictitious, open-ended experimental situation. With appropriate adaptive support (scaffolding), pre-service science teachers could be encouraged to further develop their process-oriented and results-oriented thinking patterns in this regard [162]. Considering the statements from both subtasks together, the pre-service biology teachers describe significantly more decisions in the process-related superordinate categories and also tend to describe more process-related decisions in the subcategories than the pre-service chemistry teachers. Nevertheless, subject-specific methodological (procedural) knowledge regarding scientific reasoning is rather low in the overall sample, similar to other studies in this area (e.g., [160]).

(RG3) The pre-service teachers rated diagnostics or diagnostic activities concerning experimentation as highly important for university teacher education, corresponding to the high perceived importance of diagnostics in teacher training more generally [163].

However, in-service teachers have a divergent view. Only around one-third of the in-service teachers surveyed by Lorenz [139] believed that diagnostic training is required to correctly assess students' competencies. It is possible that in-service teachers see their university studies in retrospect as focused on theoretical content and content related to their school subject, while diagnostics can only be performed when actually teaching in schools and thus can be learned in practice. Despite an increase in diagnostic knowledge during their studies compared to the beginning of their studies, pre-service teachers still rate it as below average compared to their knowledge of their subject and pedagogical knowledge [91,164]. In addition, diagnosing students' performance and difficulties in experimentation is more challenging than in other areas of science education, as it can only be assessed to a rather limited extent through written tasks. However, knowledge of the experimental competencies of the students being assessed—in particular, explicit knowledge of difficulties and misconceptions, which enables science teachers to assess student achievement against curricular expectations—is fundamental in science subjects [163]. This knowledge goes along with higher self-efficacy [124], which is in turn positively related to the successful transfer of content from teacher training (e.g., [165–167]) and ultimately to student achievement [168].

(RQ4a/b) This study's results concerning self-efficacy show that pre-service science teachers' subjective perceptions of their ability to successfully carry out general and experimentation-related diagnostic activities are significantly lower than their attitudes regarding the importance of diagnostics in teacher education. No correlations between knowledge of students' difficulties and key decisions in experimentation procedures and self-efficacy in diagnosing students' abilities in scientific reasoning acquisition were found in the sample studied here. In a study of 495 pre-service biology teachers [161], advanced students' (Master's or State Examination degree ≥ 7) self-efficacy to plan and conduct biology lessons correlated with the knowledge of what data to collect when assessing experimentation skills. No such correlation was found for pre-service teacher students at the undergraduate level. The present study examined pre-service teachers at an early stage of university teacher education, in the first third of their studies, which may be associated with a low level of knowledge in assessing self-efficacy and its associated dimensions. It is also possible that our findings were due to the surveyed students' low assessment of their own scientific inquiry competence. Conducting such surveys at the beginning of students' studies is therefore highly relevant in order to be able to promote such competences in a targeted way based on the obtained results. Khan and Krell [169] investigated the scientific reasoning competencies of pre-service Canadian science teachers and how they improved as a result of an intensive methods course including a 15-week internship. This intervention significantly improved the students' competencies in planning experiments and testing models, demonstrating that these are trainable through instruction on the scientific method including an internship with the opportunity to model, engage in, and reflect upon inquiry instruction in the science classroom. Further intervention studies demonstrate an increase in scientific reasoning competences as a result of combining different scaffolding formats [12,48,170]. In this context, it also seems interesting that, in addition to training (through scaffolds), greater background knowledge of scientific reasoning may influence pre-service teachers' competencies. Participants with prior university degrees (in other subjects) performed better in a multiple-choice questionnaire surveying scientific reasoning competencies than participants with no prior university degrees, perhaps because the former group could draw on greater background knowledge [169].

Limitations

Although the sample size was small, especially when it came to the biology and chemistry subsamples, and the findings require replication among physics education students and students at other universities, the study was able to provide insight on the views of pre-service biology and chemistry teachers in the first third of university teacher education on students' difficulties, misconceptions and key decisions in experimentation. This sample

was specifically selected because the pre-service science teachers were still at a relatively early phase of their teacher training, i.e., they were attending their first course on subject didactic training, and prior internships in their subject of study (biology/chemistry) and introductory lectures were not expected to have had much influence. The present study's sample therefore made it possible to assess the status quo of pre-service science teachers' professional knowledge and self-efficacy regarding scientific reasoning and diagnostics, in order to derive possible support measures.

At both university locations where the surveys took place, teaching education in the natural sciences focuses on scientific inquiry in research and teaching, and this context is likely to be significantly affected the students' specific content and pedagogical content knowledge. The second subject the pre-service teachers were studying might have also exerted an influence. For example, a measurement instrument for upper secondary students uncovered a distinction between "experimental ways of thinking and working" and content knowledge [170]. Pre-service teachers studying two science subjects have more opportunities to learn how to conduct scientific work, i.e., how to apply and reflect on scientific reasoning skills in different contexts and with different strategies [68], which leads to higher levels of competence [14,171].

The present survey of perceived difficulties and key decisions students face in experimentation had respondents refer to an example experimentation context (see Supplementary Material S1). While it would be possible to evaluate and carry out scientific reasoning skills in a context-free manner, it cannot be generally assumed that the same pre-service science teachers will provide the same responses on a task with the same format referring to a different context. Even though declarative knowledge of the context has only a minor influence on the successful completion of scientific reasoning test items, students with similar levels of procedural knowledge perform differently in such tasks (for an example regarding hypothesis-testing skills: [171]), which may be due to their different levels of declarative knowledge. Similarly, a different strategy can be used with each phenomenon or scientific problem to be discovered and researched, even in the same domain [68].

Another methodological limitation that deserves mention here concerns the scales measuring the students' attitudes and self-efficacy, which should be interpreted in comparison to other findings on these constructs. Neither follows the recommendation to apply scales with a large number of points [108], as they only contain four levels, potentially limiting the range of respondents' assessments. However, other studies have assessed self-efficacy with equally short Likert scales (e.g., [172]), while some studies have applied more comprehensive scales (e.g., [173]). In the future, it should be examined whether the number of scale points influences assessments of abilities and attitudes in a given area.

5. Conclusions

This study surveyed knowledge of students' difficulties, misconceptions, and key decisions in experimental problem-solving in the context of an authentic lesson plan, attitudes concerning the importance of diagnostics in teacher training, and self-efficacy related to diagnostics in experimental settings via an online questionnaire with a sample of pre-service biology and chemistry teachers. The developed instruments/tasks make it possible to survey pre-service teachers' knowledge and attitudes towards diagnosing experimentation skills within university teacher education in the subjects of biology and chemistry. They might be also transferable to physics education, and could serve as a foundation for conceptualizing university teaching-learning settings concerning scientific reasoning as well as for future intervention studies within subject didactics training regarding diagnostic knowledge of experimentation and attitudes. The results indicate that knowledge about students' misunderstandings, difficulties and problems during experimentation must be imparted to pre-service science teachers during university education. It is essential to understand scientific procedures (knowing how) and epistemic constructs (knowing why) regarding scientific reasoning in experimental problem-solving. Imparting this procedural and epistemic knowledge to teacher education at an early phase of their

studies via appropriate adaptive support (scaffolding) could help pre-service science teachers develop process-oriented and results-oriented thinking patterns as well as diagnostic skills in this regard [162]. To put this in perspective, knowledge of students' difficulties and frequent sticking points can help teachers design lessons in a student-oriented way (cf., [174]). Studies on subject-specific teaching quality in biology with respect to diagnostic competencies have shown that PCK and PK are statistically significantly related to pre-service teachers' diagnostic activities, and biology teachers' PCK is positively related to diagnostic accuracy [95]. Teachers are expected to provide students with learning opportunities that help them develop 21st-century skills, including core competencies in subject areas such as science. Zimmermann [19] concludes from her review of the literature on scientific reasoning that it is possible to teach both key features of science, i.e., the subject-specific content of scientific disciplines (e.g., biology, physics) and skills in experimentation and evaluating evidence. Previous studies on scientific reasoning show that progress has been made in research how it can be done to help students become scientifically literate adults by applying their scientific reasoning skills (cf., [19]). Therefore, it is necessary that (pre-service) science teachers are taught concepts and theories important for experimental inquiry processes, i.e., the key decisions that must be made based on concrete examples in open-ended experimentation process in order to conduct a successful experiment. In addition to knowledge about individual student learning characteristics that may be relevant to the learning process of scientific reasoning [175], procedural knowledge of scientific reasoning enables teachers to diagnose students' difficulties and misconceptions in the experimentation process. Further research using the instruments employed in this study on a larger sample of pre-service science teachers would contribute to the development of models of diagnostic competence acquisition (e.g., [176,177]) as well as professional vision (cf., [178]) for experimental problem-solving in competence-oriented science instruction and identify similarities and differences across subjects.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/educsci11100629/s1>, Task-Sheet S1: Tasks to analyse a classroom setting for experimentation—Difficulties and Approach of students.

Author Contributions: Conceptualization, D.H.-R. and M.M.; formal analysis, D.H.-R. and M.M.; investigation, D.H.-R., M.M., D.H.; writing—original draft D.H.-R.; writing—review and editing, D.H.-R., M.M., D.H., K.H.; project administration, D.H.-R., M.M.; funding acquisition, K.H., D.H.-R., M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research), at the Technische Universität Braunschweig in the project Diagonal-NaWi, grant number: 01JA1909 and at the Universität Kassel in the project PRONET², grant number: 01JA1805. Both projects are part of the “Qualitätsoffensive Lehrerbildung”, a joint initiative of the Federal Government and the Länder which aims to improve the quality of teacher training. The authors are responsible for the content of this publication. Funding of APC: The authors acknowledge support by the Open Access Publication Funds of Technische Universität Braunschweig.

Institutional Review Board Statement: All participants were students at two German universities. They took part voluntarily and signed an informed consent form. Pseudonymization of participants was guaranteed during the study and the implementation took place online-based in a stress-free environment at home. Due to all these measures in the implementation of the study, an audit by an ethics committee was waived.

Informed Consent Statement: Written informed consent was obtained from all participants involved in the study.

Data Availability Statement: Information and queries on the data used can be obtained from the authors of this article.

Acknowledgments: We would like to thank Femke Sander for her support in rating as well all pre-service teachers who participated in the study. We also thank Di Fuccia (University of Kassel) for his support in recruiting pre-service teachers with chemistry as a subject. We thank the academic

editors, and two anonymous reviewers whose recommendations substantially improved the quality of the article.

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

Appendix A

Table A1. Item parameters for the scale relevance of diagnostics in the teacher training program ($n = 50$).

Item Wording	M (SD)	r_{it}
A teacher can correctly assess students' performance in experimentation even without diagnostic components in teacher education.	2.92 (0.57)	0.55
It is important for teachers to be able to correctly assess the characteristics of the students that are relevant to learning and performance in experimentation.	3.42 (0.54)	0.51
Diagnostic skills of pre-service teachers in the assessment of experimental processes should be mandatorily promoted in teacher education.	3.30 (0.58)	0.59
Even without conducting diagnostic units in the course of study, a teacher can correctly assess the characteristics of students that are relevant to learning and performance in experimentation.	2.88 (0.52)	0.28
$M_{scale} = 3.13$; $SD_{scale} = 0.39$; Cronbach's $\alpha = 0.70$		

Appendix B

Table A2. Item analysis on domain-specific self-efficacy expectations with a pilot sample of $n = 98$.

Item Wording	Factor Loading (a_{jq})		h^2
	I	II	
Experiment-related diagnostic activities			
Even when I am under stress, I am still able to diagnose students' errors when experimenting in biology lessons.	0.818		0.697
I am confident in recognising children's specific difficulties in experimentation in biology despite great time pressure. *	0.763		0.663
In biology, I am able to identify the learning requirements of my students, even when I have little time. *	0.762		0.583
In biology, I am able to confidently understand my students' experimental skills, even when I have little time.	0.650	0.341	0.539
$n = 97$; $M_{scale} = 2.60$; $SD_{scale} = 0.44$; Cronbach's $\alpha = 0.79$			
General diagnostic activities			
In biology, I am able to integrate a diagnostic activity that accompanies learning in my teaching, even when I am under time pressure. *	0.549	0.418	0.476
Despite a high degree of heterogeneity, I am able to formulate tasks in biology with which I can appropriately test the level of knowledge of both weaker and stronger students. *		0.763	0.603
In biology, I am able to take into account the learning processes of the students when formulating individual learning objectives, even if these are very different. *		0.732	0.554
In the subject of biology, I manage to grasp the thought and work processes for gaining knowledge of my students, even when I have little time. *	0.319	0.710	0.606
$n = 95$; $M_{scale} = 2.46$; $SD_{scale} = 0.44$; Cronbach's $\alpha = 0.70$			
% of variance	46.03		12.99

Annotation: Principal component analysis with Varimax rotation; * adapted from [142].

References

- Vorholzer, A.; von Aufschnaiter, C.; Boone, W.J. Fostering Upper Secondary Students' Ability to Engage in Practices of Scientific Investigation: A Comparative Analysis of an Explicit and an Implicit Instructional Approach. *Res. Sci. Educ.* **2020**, *50*, 333–359. [CrossRef]
- NSTA/ASTE Standards for Science Teacher Preparation. 2020. Available online: <https://static.nsta.org/pdfs/2020NSTAStandards.pdf> (accessed on 2 August 2021).

3. Department for Education. Teachers' Standards. Guidance for School Leaders, School Staff and Governing Bodies. 2011. Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/665520/Teachers_Standards.pdf (accessed on 2 August 2021).
4. Beschlüsse der Kultusministerkonferenz. Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004. Available online: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Biologie.pdf (accessed on 2 August 2021).
5. Beschlüsse der Kultusministerkonferenz. Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004. Available online: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Chemie.pdf (accessed on 2 August 2021).
6. Beschlüsse der Kultusministerkonferenz. Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. Beschluss vom 16.12.2004. Available online: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Physik-Mittleren-SA.pdf (accessed on 2 August 2021).
7. Wellnitz, N.; Hecht, M.; Heitmann, P.; Kauertz, A.; Mayer, J.; Sumfleth, E.; Walpuski, M. Modellierung des Kompetenzteilbereichs Naturwissenschaftliche Untersuchungen. *Z. Erzieh.* **2017**, *20*, 556–584. [CrossRef]
8. Darling-Hammond, L. Teacher Quality and Student Achievement. *Educ. Policy Anal. Arch.* **2000**, *8*, 1. [CrossRef]
9. Morrell, P.D.; Park Rogers, M.A.; Pyle, E.J.; Roehrig, G.; Veal, W.R. Preparing Teachers of Science for 2020 and Beyond: Highlighting Changes to the NSTA/ASTE Standards for Science Teacher Preparation. *J. Sci. Teach. Educ.* **2020**, *31*, 1–7. [CrossRef]
10. Committee on the Development of an Addendum to the National Science Education Standards on Scientific Inquiry; Board on Science Education; Division of Behavioral and Social Sciences and Education; National Research Council. *Inquiry and the National Science Education Standards: A Guide for Teaching and Learning*; Olson, S., Loucks-Horsley, S., Eds.; National Academies Press: Washington, DC, USA, 2000; p. 9596, ISBN 9780309064767.
11. Kultusministerkonferenz. Ländergemeinsame Inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung. Beschluss der Kultusministerkonferenz vom 16.10.2008 i. d. F. vom 16.05.2019. Available online: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf (accessed on 2 August 2021).
12. Bruckermann, T.; Aschermann, E.; Bresges, A.; Schlüter, K. Metacognitive and Multimedia Support of Experiments in Inquiry Learning for Science Teacher Preparation. *Int. J. Sci. Educ.* **2017**, *39*, 701–722. [CrossRef]
13. Hodson, D. Learning Science, Learning about Science, Doing Science: Different Goals Demand Different Learning Methods. *Int. J. Sci. Educ.* **2014**, *36*, 2534–2553. [CrossRef]
14. Hartmann, S.; Upmeier zu Belzen, A.; Krüger, D.; Pant, H.A. Scientific Reasoning in Higher Education: Constructing and Evaluating the Criterion-Related Validity of an Assessment of Preservice Science Teachers' Competencies. *Z. Psychol.* **2015**, *223*, 47–53. [CrossRef]
15. Mathesius, S.; Hartmann, S.; Upmeier zu Belzen, A.; Krüger, D. Scientific Reasoning as an Aspect of Preservice Biology Teacher Education. Assessing competencies using a paper-pencil test. In *The Future of Biology Education Research*; Tal, T., Yarden, A., Eds.; The Technion, Israel Institute of Technology/The Weizmann Institute of Science: Haifa, Israel, 2016; pp. 93–110.
16. Bernholt, S.; Neumann, K.; Nentwig, P. *Making It Tangible: Learning Outcomes in Science Education*; Waxmann: Münster, Germany, 2012; pp. 13–28.
17. Opitz, A.; Heene, M.; Fischer, F. Measuring Scientific Reasoning—A Review of Test Instruments. *Educ. Res. Eval.* **2017**, *23*, 78–101. [CrossRef]
18. Koslowski, B. *Theory and Evidence: The Development of Scientific Reasoning*; MIT Press: Cambridge, MA, USA, 1996.
19. Zimmerman, C. The Development of Scientific Reasoning Skills. *Dev. Rev.* **2000**, *20*, 99–149. [CrossRef]
20. Paul, R.; Elder, L. *The Thinker's Guide to Scientific Thinking: Based on Critical Thinking Concepts & Principles*; Thinker's Guide Library: Tomales, CA, USA, 2012.
21. Kuhn, D.; Amsel, E.; O'Loughlin, M. The Development of Scientific Thinking Skills. In *Developmental Psychology Series*; Academic Press: San Diego, CA, USA, 1988.
22. Klahr, D. Exploring Science. In *The Cognition and Development of Discovery Processes*; MIT Press: Cambridge, MA, USA, 2000.
23. Schafersman, S.D. An Introduction to Science: Scientific Thinking and the Scientific Method. Available online: <http://www.geo.sunysb.edu/esp/files/scientific-method.html> (accessed on 3 August 2021).
24. Janoušková, S.; Pyskatá Rathouská, L.; Žák, V.; Urválková, E.S. The Scientific Thinking and Reasoning Framework and Its Applicability to Manufacturing and Services Firms in Natural Sciences. *Res. Sci. Technol. Educ.* **2021**, 1–22. [CrossRef]
25. Dowd, J.E.; Thompson, R.J.; Schiff, L.A.; Reynolds, J.A. Understanding the Complex Relationship between Critical Thinking and Science Reasoning among Undergraduate Thesis Writers. *Life Sci. Educ.* **2018**, *17*, ar4. [CrossRef]
26. Kuhn, D. *The Skills of Argument*; Cambridge University Press: Cambridge, UK, 1991.
27. Fischer, F.; Kollar, I.; Ufer, S.; Sodian, B.; Hussmann, H.; Pekrun, R.; Neuhaus, B.; Dorner, B.; Pankofer, S.; Fischer, M.; et al. Scientific Reasoning and Argumentation: Advancing an Interdisciplinary Research Agenda in Education. *Frontline Learn. Res.* **2014**, *2*, 28–45. [CrossRef]
28. Dunbar, K.; Fugelsang, J. Scientific Thinking and Reasoning. In *The Cambridge Handbook of Thinking and Reasoning*; Holyoak, K.J., Morrison, R.G., Eds.; Cambridge University Press: Cambridge, UK, 2005; pp. 705–725.

29. Krell, M.; Redman, C.; Mathesius, S.; Krüger, D.; van Driel, J. Assessing Pre-Service Science Teachers' Scientific Reasoning Competencies. *Res. Sci. Educ.* **2018**, *50*, 2305–2329. [CrossRef]
30. Lawson, A.E. The Nature and Development of Scientific Reasoning: A Synthetic View. *Int. J. Sci. Math. Educ.* **2004**, *2*, 307–338. [CrossRef]
31. Morris, B.J.; Croker, S.; Masnick, A.; Zimmerm, C. The Emergence of Scientific Reasoning. In *Current Topics in Children's Learning and Cognition*; Kloos, H., Ed.; IntechOpen: London, UK, 2012; pp. 61–82.
32. Kind, P.; Osborne, J. Styles of Scientific Reasoning: A Cultural Rationale for Science Education? *Sci. Ed.* **2017**, *101*, 8–31. [CrossRef]
33. Arnold, J.C.; Mühlhng, A.; Kremer, K. Exploring Core Ideas of Procedural Understanding in Scientific Inquiry Using Educational Data Mining. *Res. Sci. Technol. Educ.* **2021**, 1–21. [CrossRef]
34. Mayer, J. Erkenntnisgewinnung als Wissenschaftliches Problemlösen. In *Theorien in der Biologiedidaktischen Forschung: Ein Handbuch für Lehramtsstudenten und Doktoranden*; Krüger, D., Vogt, H., Eds.; Springer: Berlin, Germany; New York, NY, USA, 2007; pp. 177–186.
35. Baur, A.; Emden, M. How to Open Inquiry Teaching? An Alternative Teaching Scaffold to Foster Students' Inquiry Skills. *Chem. Teach. Int.* **2021**, *3*, 20190013. [CrossRef]
36. Rieß, W.; Wirtz, M.A.; Barzel, B.; Schulz, A. Integration der theoretischen und empirischen Befunde zum Experimentieren im mathematisch-naturwissenschaftlichen Unterricht. In *Experimentieren im Mathematisch-Naturwissenschaftlichen Unterricht: Schüler*; Waxmann Verlag: Münster, Germany, 2012; pp. 353–364.
37. Andersen, C.; Garcia-Mila, M. Scientific Reasoning During Inquiry. In *Science Education*; Taber, K.S., Akpan, B., Eds.; SensePublishers: Rotterdam, The Netherlands, 2017; pp. 105–117.
38. Arnold, J.C.; Boone, W.J.; Kremer, K.; Mayer, J. Assessment of Competencies in Scientific Inquiry through the Application of Rasch Measurement Techniques. *Educ. Sci.* **2018**, *8*, 184. [CrossRef]
39. Meier, M. *Entwicklung und Prüfung eines Instrumentes zur Diagnose der Experimentierkompetenz von Schülerinnen und Schülern*; Logos Verlag: Berlin, Germany, 2016.
40. Hammann, M.; Phan, T.T.H.; Bayrhuber, H. Experimentieren als Problemlösen: Lässt sich das SDDS-Modell Nutzen, um Unterschiedliche Dimensionen beim Experimentieren zu messen? In *Kompetenzdiagnostik*; Prenzel, M., Gogolin, I., Krüger, H.-H., Eds.; Springer-Verlag GmbH: Wiesbaden, Germany, 2007; pp. 33–50.
41. Nowak, K.H.; Nehring, A.; Tiemann, R.; Upmeier zu Belzen, A. Assessing Students' Abilities in Processes of Scientific Inquiry in Biology Using a Paper-and-Pencil Test. *J. Biol. Educ.* **2013**, *47*, 182–188. [CrossRef]
42. Nehring, A.; Nowak, K.H.; zu Belzen, A.U.; Tiemann, R. Predicting Students' Skills in the Context of Scientific Inquiry with Cognitive, Motivational, and Sociodemographic Variables. *Int. J. Sci. Educ.* **2015**, *37*, 1343–1363. [CrossRef]
43. Nehring, A.; Stiller, J.; Nowak, K.H.; Upmeier zu Belzen, A.; Tiemann, R. Naturwissenschaftliche Denk- und Arbeitsweisen im Chemieunterricht-eine Modellbasierte Videostudie zu Lerngelegenheiten für den Kompetenzbereich der Erkenntnisgewinnung. *ZfDN* **2016**, *22*, 77–96. [CrossRef]
44. Kraeva, L. *Problemlösestrategien von Schülerinnen und Schülern Diagnostizieren*; Logos Verlag: Berlin, Germany, 2020.
45. Neumann, I. *Beyond Physics Content Knowledge: Modeling Competence Regarding Nature of Scientific Inquiry and Nature of Scientific Knowledge*; Logos Verlag: Berlin, Germany, 2011.
46. Maiseyenko, V.; Schecker, H.; Nawrath, D. Kompetenzorientierung des Naturwissenschaftlichen Unterrichts-Symbiotische Kooperation bei der Entwicklung eines Modells Experimenteller Kompetenz. *PhyDid A Phys. Didakt. Sch. Hochsch.* **2013**, *1*, 1–17.
47. Nababan, N.P.; Nasution, D.; Jayanti, R.D. The Effect of Scientific Inquiry Learning Model and Scientific Argumentation on The Students' Science Process Skill. *J. Phys. Conf. Ser.* **2019**, *1155*, 012064. [CrossRef]
48. Hilfert-Rüppell, D.; Eghtessad, A.; Höner, K. Interaktive Videovignetten aus naturwissenschaftlichem Unterricht—Förderung der Diagnosekompetenz von Lehramtsstudierenden hinsichtlich der Experimentierfähigkeit von Schülerinnen und Schülern. *Medien Pädagogik* **2018**, *31*, 125–142. [CrossRef]
49. Alfieri, L.; Brooks, P.J.; Aldrich, N.J.; Tenenbaum, H.R. Does Discovery-Based Instruction Enhance Learning? *J. Educ. Psychol.* **2011**, *103*, 1–18. [CrossRef]
50. Abrams, E.; Southerland, S.A.; Evans, C. Inquiry in the Classroom: Identifying Necessary Components of a Useful Definition. In *Inquiry in the Science Classroom: Challenges and Opportunities*; Abrams, E., Southerland, S., Silva, P., Eds.; Information Age Publishing: Charlotte, NC, USA, 2008; pp. 11–42.
51. Barzel, B.; Reinhoffer, B.; Schrenk, B. Das Experimentieren im Unterricht. In *Experimentieren im Mathematisch-Naturwissenschaftlichen Unterricht: Schüler Lernen Wissenschaftlich Denken und Arbeiten*; Rieß, W., Wirtz, M.A., Barzel, B., Schulz, A., Eds.; Waxmann Verlag: Münster, Germany, 2012; pp. 103–128.
52. Ural, E. The Effect of Guided-Inquiry Laboratory Experiments on Science Education Students' Chemistry Laboratory Attitudes, Anxiety and Achievement. *J. Educ. Train. Stud.* **2016**, *4*, 217–227. [CrossRef]
53. Hmelo-Silver, C.E.; Duncan, R.G.; Chinn, C.A. Scaffolding and Achievement in Problem-Based and Inquiry Learning: A Response to Kirschner, Sweller, and Clark (2006). *Educ. Psychol.* **2007**, *42*, 99–107. [CrossRef]
54. Vorholzer, A.; von Aufschnaiter, C. Guidance in Inquiry-Based Instruction-An Attempt to Disentangle a Manifold Construct. *Int. J. Sci. Educ.* **2019**, *41*, 1562–1577. [CrossRef]
55. Höner, K.; Eghtessad, A.; Hilfert-Rüppell, D.; Kraeva, L. Naturwissenschaftliches Potenzial?—Diagnose von Schülerfähigkeiten zum experimentellen Problemlösen. *J. Für Begabtenförderung* **2017**, *2*, 8–23.

56. Hammann, M.; Phan, T.T.H.; Ehmer, M.; Grimm, T. Assessing Pupils' Skills in Experimentation. *J. Biol. Exp.* **2008**, *42*, 66–72. [CrossRef]
57. National Academy of Engineering and National Research Council. *STEM Integration in K-12 Education: Status, Prospects, and an Agenda for Research*; National Academies Press: Washington, DC, USA, 2014.
58. Bell, R.L.; Blair, L.M.; Crawford, B.A.; Lederman, N.G. Just Do It? Impact of a Science Apprenticeship Program on High School Students' Understandings of the Nature of Science and Scientific Inquiry. *J. Res. Sci. Teach.* **2003**, *40*, 487–509. [CrossRef]
59. Klahr, D.; Nigam, M. The Equivalence of Learning Paths in Early Science Instruction: Effect of Direct Instruction and Discovery Learning. *Psychol. Sci.* **2004**, *15*, 661–667. [CrossRef]
60. Kirschner, P.A.; Sweller, J.; Clark, R.E. Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educ. Psychol.* **2006**, *41*, 75–86. [CrossRef]
61. Dunbar, K.; Klahr, D. Developmental Differences in Scientific Discovery Processes. In *Complex Information Processing: The Impact of Herbert, A. Simon*; Lawrence Erlbaum Associates, Inc.: New Jersey, NJ, USA, 1989; pp. 109–143.
62. Chinn, C.A.; Brewer, W.F. The Role of Anomalous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction. *Rev. Educ. Res.* **1993**, *63*, 1–49. [CrossRef]
63. Hammann, M.; Phan, T.T.H.; Ehmer, M.; Bayrhuber, H. Fehlerfrei Experimentieren. *Math. Nat. Unterr.* **2006**, *59*, 292–299.
64. Klahr, D.; Fay, A.L.; Dunbar, K. Heuristics for Scientific Experimentation: A Developmental Study. *Cogn. Psychol.* **1993**, *25*, 111–146. [CrossRef]
65. Lubben, F.; Millar, R. Children's Ideas about the Reliability of Experimental Data. *Int. J. Sci. Educ.* **1996**, *18*, 955–968. [CrossRef]
66. Baur, A. Fehler, Fehlkonzepte und Spezifische Vorgehensweisen von Schülerinnen und Schülern beim Experimentieren: Ergebnisse einer Videogestützten Beobachtung. *ZfDN* **2018**, *24*, 115–129. [CrossRef]
67. Hilfert-Rüppell, D.; Höner, K. *Diagnosing Student's Scientific Reasoning Skills Based on Authentic Videos of Inquiry Experiments (Working Title)*; TU Braunschweig: Braunschweig, Germany, 2021.
68. Glaser, R.; Schauble, L.; Raghavan, K.; Zeitz, C. Scientific Reasoning Across Different Domains. In *Computer-Based Learning Environments and Problem Solving*; De Corte, E., Linn, M.C., Mandl, H., Verschaffel, L., Eds.; Springer: Berlin/Heidelberg, Germany, 1992; pp. 345–371.
69. Germann, P.J.; Aram, R.; Burke, G. Identifying Patterns and Relationships among the Responses of Seventh-Grade Students to the Science Process Skill of Designing Experiments. *J. Res. Sci. Teach.* **1996**, *33*, 79–99. [CrossRef]
70. Chen, Z.; Klahr, D. All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. *Child Dev.* **1999**, *70*, 1098–1120. [CrossRef] [PubMed]
71. Kirchner, S.; Priemer, B. Probleme von Schülern Mit Offenen Experimentieraufgaben in Physik. In *Naturwissenschaftlicher Unterricht im internationalen Vergleich*; LIT Verlag: Münster, Germany, 2007; pp. 346–348.
72. Wahser, I.; Sumfleth, E. Training Experimenteller Arbeitsweisen Zur Unterstützung Kooperativer Kleingruppenarbeit Im Fach Chemie. *ZfDN* **2008**, *14*, 219–241.
73. Meier, M.; Mayer, J. Experimentierkompetenz praktisch erfassen—Entwicklung und Validierung eines anwendungsbezogenen Aufgabendesigns. In *Lehr- und Lernforschung in der Biologiedidaktik*; Harms, U., Bogner, F.X., Eds.; StudienVerlag: Innsbruck, Austria, 2012; Volume 5, pp. 81–98.
74. Meier, M.; Mayer, J. Selbständiges Experimentieren: Entwicklung und Einsatz eines anwendungsbezogenen Aufgabendesigns. *MNU* **2014**, *67*, 4–10.
75. Kechel, J.-H.; Wodzinski, R. Schülerschwierigkeiten beim Experimentieren zum Hooke'schen Gesetz. In *Authentizität und Lernen-das Fach in der Fachdidaktik*; Universität Regensburg: Regensburg, Germany, 2016; pp. 170–172.
76. Nerdel, C. Kompetenzorientiert und aufgabenbasiert für Schule und Hochschule. In *Grundlagen der Naturwissenschaftsdidaktik*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 2017.
77. Carey, S.; Evans, R.; Honda, M.; Jay, E.; Unger, C. 'An Experiment Is When You Try It and See If It Works': A Study of Grade 7 Students' Understanding of the Construction of Scientific Knowledge. *Int. J. Sci. Educ.* **1989**, *11*, 514–529. [CrossRef]
78. Schauble, L.; Klopfer, L.E.; Raghavan, K. Students' Transition from an Engineering Model to a Science Model of Experimentation. *J. Res. Sci. Teach.* **1991**, *28*, 859–882. [CrossRef]
79. Chinn, C.A.; Brewer, W.F. An Empirical Test of a Taxonomy of Responses to Anomalous Data in Science. *J. Res. Sci. Teach.* **1998**, *35*, 623–654. [CrossRef]
80. Ludwig, T.; Priemer, B. Begründungen und Überzeugungen beim Beibehalten und Verwerfen von eigenen Hypothesen in Real- und Simulationsexperimenten. In *Konzepte Fachdidaktischer Strukturierung für den Unterricht*; Bernholt, S., Ed.; LIT Verlag: Münster, Germany, 2012; pp. 313–315.
81. De Jong, T.; Van Joolingen, W.R. Scientific Discovery Learning with Computer Simulations of Conceptual Domains. *Rev. Educ. Res.* **1998**, *68*, 179–201. [CrossRef]
82. Arnold, J.; Kremer, K.; Mayer, J. Understandig Students' Experiments—What Kind of Support Do They Need in Inquiry Tasks? *Int. J. Sci. Educ.* **2014**, *36*, 2719–2749. [CrossRef]
83. Arnold, J.; Kremer, K.; Mayer, J. Schüler als Forscher. Experimentieren kompetenzorientiert Unterrichten und Beurteilen. *MNU* **2014**, *67*, 83–91.
84. Hammann, M. Kompetenzentwicklungsmodelle: Merkmale und ihre Bedeutung - dargestellt anhand von Kompetenzen beim Experimentieren. *MNU* **2004**, *57*, 196–203.

85. Dunbar, K. Concept Discovery in a Scientific Domain. *Cogn. Sci.* **1993**, *17*, 397–434. [CrossRef]
86. Härtig, H.; Neumann, K.; Erb, R. Experimentieren als Interaktion von Situation und Person: Ergebnisse einer Expertenbefragung. *ZfjDN* **2017**, *23*, 71–80. [CrossRef]
87. Cappell, J. *Fachspezifische Diagnosekompetenz Angehender Physiklehrkräfte in der Ersten Ausbildungsphase*; Studien zum Physik- und Chemielernen; Logos Verlag: Berlin, Germany, 2013.
88. Bögeholz, S.; Joachim, C.; Hasse, S.; Hammann, M. Kompetenzen von (angehenden) Biologielehrkräften Zur Beurteilung von Experimentierkompetenzen. *Unterrichtswissenschaft* **2016**, *44*, 40–54.
89. Draude, M. *Die Kompetenz von Physiklehrkräften, Schwierigkeiten von Schülerinnen und Schülern beim Eigenständigen Experimentieren zu Diagnostizieren*; Logos Verlag Berlin: Berlin, Germany, 2016.
90. Kechel, J.-H. *Schülerschwierigkeiten beim Eigenständigen Experimentieren: Eine Qualitative Studie am Beispiel einer Experimentieraufgabe zum Hooke'schen Gesetz*; Logos Verlag Berlin: Berlin, Germany, 2016.
91. Dübbelde, G. *Diagnostische Kompetenzen Angehender Biologie-Lehrkräfte im Bereich der Naturwissenschaftlichen Erkenntnisgewinnung*; Universität Kassel: Kassel, Germany, 2013. Available online: <https://kobra.uni-kassel.de/handle/123456789/2013122044701> (accessed on 3 August 2021).
92. Beretz, A.-K.; Lengnink, K.; Aufschnaiter, C. Diagnostische Kompetenz gezielt Fördern-Videoeinsatz im Lehramtsstudium Mathematik und Physik. In *Diagnose und Förderung Heterogener Lerngruppen: Theorien, Konzepte und Beispiele aus der MINT-Lehrerbildung*; Selter, C., Michaelis, J., Lengnink, K., Knipping, C., Hößle, C., Hußmann, S., Eds.; Waxmann: Münster, Germany; New York, NY, USA, 2017; pp. 149–168.
93. Schrader, J. Theorie und Praxis der Erwachsenenbildung. In *Struktur und Wandel der Weiterbildung*; Bertelsmann: Bielefeld, Germany, 2011.
94. Chernikova, O.; Heitzmann, N.; Fink, M.C.; Timothy, V.; Seidel, T.; Fischer, F. Facilitating Diagnostic Competences in Higher Education-A Meta-Analysis in Medical and Teacher Education. *Educ. Psychol. Rev.* **2020**, *32*, 157–196. [CrossRef]
95. Kramer, M.; Förtsch, C.; Boone, W.J.; Seidel, T.; Neuhaus, B.J. Investigating Pre-Service Biology Teachers' Diagnostic Competences: Relationships between Professional Knowledge, Diagnostic Activities, and Diagnostic Accuracy. *Educ. Sci.* **2021**, *11*, 89. [CrossRef]
96. Shulman, L.S. Knowledge and Teaching. Foundations of the New Reform. *Harv. Educ. Rev.* **1987**, *57*, 1–22. [CrossRef]
97. Kunter, M.; Klusmann, U.; Baumert, J. Professionelle Kompetenz von Mathematiklehrkräften: Das COACTIV-Modell. In *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung*; Zlatkin-Troitschanskaia, O., Beck, K., Sembill, D., Nickolaus, R., Mulder, R., Eds.; Beltz: Weinheim, Germany, 2009; pp. 153–165.
98. Barth, V.L.; Piwowar, V.; Kumschick, I.R.; Ophardt, D.; Thiel, F. The Impact of Direct Instruction in a Problem-Based Learning Setting. Effects of a Video-Based Training Program to Foster Preservice Teachers' Professional Vision of Critical Incidents in the Classroom. *Int. J. Educ. Res.* **2019**, *95*, 1–12. [CrossRef]
99. Kramer, M.; Förtsch, C.; Stürmer, J.; Förtsch, S.; Seidel, T.; Neuhaus, B.J. Measuring Biology Teachers' Professional Vision: Development and Validation of a Video-Based Assessment Tool. *Cogent. Educ.* **2020**, *7*, 1823155. [CrossRef]
100. Kramer, M.; Förtsch, C.; Neuhaus, B.J. Can Pre-Service Biology Teachers' Professional Knowledge and Diagnostic Activities Be Fostered by Self-Directed Knowledge Acquisition via Texts? *Educ. Sci.* **2021**, *11*, 244. [CrossRef]
101. Blömeke, S.; Gustafsson, J.-E.; Shavelson, R.J. Beyond Dichotomies: Competence Viewed as a Continuum. *Z. Psychol.* **2015**, *223*, 3–13. [CrossRef]
102. Seidel, T.; Stürmer, K. Modeling and Measuring the Structure of Professional Vision in Preservice Teachers. *Am. Educ. Res. J.* **2014**, *51*, 739–771. [CrossRef]
103. Blomberg, G.; Sherin, M.G.; Renkl, A.; Glogger, I.; Seidel, T. Understanding Video as a Tool for Teacher Education: Investigating Instructional Strategies to Promote Reflection. *Instr. Sci.* **2014**, *42*, 443–463. [CrossRef]
104. Meissel, K.; Meyer, F.; Yao, E.S.; Rubie-Davies, C.M. Subjectivity of Teacher Judgments: Exploring Student Characteristics That Influence Teacher Judgments of Student Ability. *Teach. Teach. Educ.* **2017**, *65*, 48–60. [CrossRef]
105. Baumert, J.; Kunter, M. Das Kompetenzmodell von COACTIV. In *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV*; Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., Neubrand, M., Eds.; Waxmann: Münster, Germany, 2011; pp. 29–53.
106. Bandura, A. *Social Foundations of Thought and Action: A Social Cognitive Theory*; Prentice-Hall, Inc.: Englewood Cliffs, NJ, USA, 1986.
107. Bong, M.; Skaalvik, E.M. Academic Self-Concept and Self-Efficacy: How Different Are They Really? *Educ. Psychol. Rev.* **2003**, *15*, 1–40. [CrossRef]
108. Bandura, A.; Freeman, W.H.; Lightsey, R. Self-Efficacy: The Exercise of Control. *J. Cogn. Psychother.* **1999**, *13*, 158–166. [CrossRef]
109. Bandura, A. Guide for Constructing Self-Efficacy Scales. In *Self-Efficacy Beliefs of Adolescents*; Pajares, F., Urdan, T., Eds.; Information Age Publishing: Charlotte, NC, USA, 2006; Volume 5, pp. 307–337.
110. Pajares, F. Self-Efficacy Beliefs in Academic Settings. *Rev. Educ. Res.* **1996**, *66*, 543–578. [CrossRef]
111. Andrew, S. Self-Efficacy as a Predictor of Academic Performance in Science. *J. Adv. Nurs.* **1998**, *27*, 596–603. [CrossRef] [PubMed]
112. Schunk, D.H.; Pajares, F. Self-Efficacy Theory. In *Handbook of Motivation at School*; Wenzel, K.R., Wigfield, A., Eds.; Routledge/Taylor & Francis Group: New York, NY, USA, 2009; pp. 35–53.
113. Komarraju, M.; Nadler, D. Self-Efficacy and Academic Achievement: Why Do Implicit Beliefs, Goals, and Effort Regulation Matter? *Learn. Individ. Differ.* **2013**, *25*, 67–72. [CrossRef]

114. Handtke, K.; Bögeholz, S. Self-Efficacy Beliefs of Interdisciplinary Science Teaching (Self-ST) Instrument: Drafting a Theory-Based Measurement. *Educ. Sci.* **2019**, *9*, 247. [CrossRef]
115. Hackett, G.; Betz, N.E. An Exploration of the Mathematics Self-Efficacy/Mathematics Performance Correspondence. *J. Res. Math. Educ.* **1989**, *20*, 261–273. [CrossRef]
116. Pajares, F.; Kranzler, J. Self-Efficacy Beliefs and General Mental Ability in Mathematical Problem-Solving. *Contemp. Educ. Psychol.* **1995**, *20*, 426–443. [CrossRef]
117. Pajares, F.; Graham, L. Self-Efficacy, Motivation Constructs, and Mathematics Performance of Entering Middle School Students. *Contemp. Educ. Psychol.* **1999**, *24*, 124–139. [CrossRef]
118. Fencil, H.; Scheel, K. Engaging Students: An Examination of the Effects of Teaching Strategies on Self-Efficacy and Course Climate in a Nonmajors Physics Course. *J. Coll. Sci. Teach.* **2005**, *35*, 20.
119. Dalgety, J.; Coll, R.K. Exploring First-Year Science Students' Chemistry Self-Efficacy. *Int. J. Sci. Math. Educ.* **2006**, *4*, 97–116. [CrossRef]
120. Yürük, N. The Predicators of Pre-Service Elementary Teachers' Anxiety about Teaching Science. *J. Balt. Sci. Educ.* **2011**, *10*, 17–26.
121. Kurbanoglu, N.I.; Akim, A. The Relationships between University Students' Chemistry Laboratory Anxiety, Attitudes, and Self-Efficacy Beliefs. *Aust. J. Teach. Educ.* **2010**, *35*, 48–59. [CrossRef]
122. Trujillo, G.; Tanner, K.D. Considering the Role of Affect in Learning: Monitoring Students' Self-Efficacy, Sense of Belonging, and Science Identity. *Life Sci. Educ.* **2014**, *13*, 6–15. [CrossRef] [PubMed]
123. Riese, J.; Reinhold, P. Empirische Erkenntnisse zur Struktur Professioneller Handlungskompetenz von angehenden Physik-lehrkräften. *ZfDN* **2010**, *16*, 167–187.
124. Mahler, D.; Großschedl, J.; Harms, U. Opportunities to Learn for Teachers' Self-Efficacy and Enthusiasm. *Educ. Res. Int.* **2017**, *2017*, 1–17. [CrossRef]
125. Klug, J.; Bruder, S.; Schmitz, B. Which Variables Predict Teachers Diagnostic Competence When Diagnosing Students' Learning Behavior at Different Stages of a Teacher's Career? *Teach. Teach.* **2016**, *22*, 461–484. [CrossRef]
126. Holzberger, D.; Philipp, A.; Kunter, M. How Teachers' Self-Efficacy Is Related to Instructional Quality: A Longitudinal Analysis. *J. Educ. Psychol.* **2013**, *105*, 774–786. [CrossRef]
127. Blömeke, S. Qualitativ-Quantitativ, Induktiv-Deduktiv, Prozess-Produkt, National-International. In *Forschung zur Lehrerbildung*; Lüders, M., Ed.; Waxmann Verlag: Münster, Germany, 2007; pp. 13–36.
128. Barke, H.-D.; Hazari, A.; Yitbarek, S. *Misconceptions in Chemistry: Addressing Perceptions in Chemical Education*; Springer: Berlin/Heidelberg, Germany, 2009.
129. Tolsdorf, Y.; Markic, S. Exploring Chemistry Student Teachers' Diagnostic Competence-A Qualitative Cross-Level Study. *Educ. Sci.* **2017**, *7*, 86. [CrossRef]
130. Tolsdorf, Y.; Markic, S. Development of an Instrument and Evaluation Pattern for the Analysis of Chemistry Student Teachers' Diagnostic Competence. *Int. J. Phys. Chem. Educ.* **2017**, *9*, 1–10.
131. Capizzi, A.M.; Fuchs, L.S. Effects of Curriculum-Based Measurement with and without Diagnostic Feedback on Teacher Planning. *Remedial Spec. Educ.* **2005**, *26*, 159–174. [CrossRef]
132. Cramer, C. Beurteilung Des Bildungswissenschaftlichen Studiums Durch Lehramtsstudierende in Der Ersten Ausbildungsphase Im Längsschnitt. *Z. Pädagogik* **2013**, *59*, 66–82.
133. Bicak, B.E.; Borchert, C.E.; Höner, K. Measuring and Fostering Preservice Chemistry Teachers' Scientific Reasoning Competency. *Educ. Sci.* **2021**, *11*, 496. [CrossRef]
134. Osborne, J. The 21st Century Challenge for Science Education: Assessing Scientific Reasoning. *Think. Ski. Creat.* **2013**, *10*, 265–279. [CrossRef]
135. Shavelson, R.J. Assessing Business-Planning Competence using the Collegiate Learning Assessment as a Prototype. *Empir. Res. Vocat. Educ. Train.* **2012**, *4*, 77–90. [CrossRef]
136. Roberts, R.; Gott, R.; Glaesser, J. Students' approaches to open-ended science investigation: The importance of substantive and procedural understanding. *Res. Pap. Educ.* **2010**, *25*, 377–407. [CrossRef]
137. Mathesius, S.; Upmeier zu Belzen, A.; Krüger, D. Eyetracking als Methode zur Untersuchung von Lösungsprozessen bei Multiple-Choice-Aufgaben zum wissenschaftlichen Denken. In *Lehr- und Lernforschung in der Biologiedidaktik*; Hammann, M., Lindner, M., Eds.; Studienverlag: Innsbruck, Germany, 2018; pp. 225–244.
138. Osbeck, L.M. Scientific Reasoning as Sense-Making: Implications for Qualitative Inquiry. *Qual. Psychol.* **2014**, *1*, 34–46. [CrossRef]
139. Lorenz, C. Schriften aus der Fakultät Humanwissenschaften der Otto-Friedrich-Universität Bamberg. In *Diagnostische Kompetenz von Grundschullehrkräften: Strukturelle Aspekte und Bedingungen*; University of Bamberg Press: Bamberg, Germany, 2011.
140. Cronbach, L.J.; Meehl, P.E. Construct Validity in Psychological Tests. *Psychol. Bull.* **1955**, *52*, 281–302. [CrossRef]
141. Schwarzer, R.; Jerusalem, M. Das Konzept Der Selbstwirksamkeit. In *Selbstwirksamkeit und Motivationsprozesse in Bildungsinstitutionen*; Jerusalem, M., Hopf, D., Eds.; Beltz Verlag: Weinheim, Germany; Basel, Switzerland, 2002; pp. 28–53.
142. Sprenger, M.; Wartha, S.; Lipowsky, F. *Skalenhandbuch der Fortbildungsstudie QUASUM-Wirkungen einer Qualifizierungsmaßnahme zum Thema Rechenstörungen auf das Diagnostische Wissen und die Selbstwirksamkeitserwartungen von Mathematiklehrpersonen*; Pädagogische Hochschule Karlsruhe: Karlsruhe, Germany, 2015; Unpublished work.
143. Mayring, P. *Qualitative Inhaltsanalyse: Grundlagen und Techniken*; Beltz: Weinheim, Germany; Basel, Switzerland, 2015.

144. Stamann, C.; Janssen, M.; Schreier, M. Qualitative Inhaltsanalyse-Versuch einer Begriffsbestimmung und Systematisierung. *Forum Qual. Soc. Res.* **2016**, *17*, 16. [CrossRef]
145. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
146. Taber, K. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Res. Sci. Educ.* **2017**, *48*, 1273–1296. [CrossRef]
147. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1988.
148. Fensham, P. Science for All. *Educ. Leadersh.* **1986**, *44*, 18–23.
149. Next Generation Science Standards (NGSS). Topic Arrangements of the Next Generation Science Standards. Available online: <https://www.nextgenscience.org/sites/default/files/AllTopic.pdf> (accessed on 4 August 2021).
150. Hessisches Kultusministerium. Kerncurriculum Gymnasiale Oberstufe. Biologie. Available online: <https://kultusministerium.hessen.de/sites/default/files/media/kcgo-bio.pdf> (accessed on 5 August 2021).
151. Hessisches Kultusministerium. Kerncurriculum Gymnasiale Oberstufe. Chemie. Available online: <https://kultusministerium.hessen.de/sites/default/files/media/kcgo-ch.pdf> (accessed on 5 August 2021).
152. Niedersächsisches Kultusministerium. Kerncurriculum für das Gymnasium-Gymnasiale Oberstufe, die Gesamtschule-Gymnasiale Oberstufe, das Berufliche Gymnasium das Abendgymnasium, das Kolleg. Biologie. Available online: https://cuvo.nibis.de/cuvo.php?p=search&k0_0=Dokumentenart&v0_0=Kerncurriculum&k0_1=Schulbereich&v0_1=Sek+II&k0_2=Fach&v0_2=Biologie& (accessed on 5 August 2021).
153. Niedersächsisches Kultusministerium. Kerncurriculum für das Gymnasium-Gymnasiale Oberstufe, die Gesamtschule-Gymnasiale Oberstufe, das Berufliche Gymnasium das Abendgymnasium, das Kolleg. Chemie. Available online: https://cuvo.nibis.de/cuvo.php?p=search&k0_0=Dokumentenart&v0_0=Kerncurriculum&k0_1=Schulbereich&v0_1=Sek+II&k0_2=Fach&v0_2=Chemie& (accessed on 5 August 2021).
154. Ziepprecht, K.; Meier, M. Umsetzung und Weiterentwicklung von Modellen zur curricularen Vernetzung in hochschuldidaktischen Lernumgebungen in PRONET und PRONET². In *Vielfältige Wege biologiedidaktischer Forschung. Vom Lernort Natur zur Naturwissenschaftlichen Erkenntnisgewinnung in die Lehrerprofessionalisierung*; Meier, M., Wulff, C., Ziepprecht, K., Eds.; Waxmann: Münster, Germany, 2021; pp. 203–218.
155. Becerra-Labra, C.; Gras-Martí, A.; Martínez Torregrosa, J. Teaching Physics with a Fundamental-Problem-Based Approach: Effects on Conceptual Learning, Attitudes and Interests of University Students. *Rev. Bras. Ensino Fis.* **2006**, *29*, 95–103. [CrossRef]
156. Krüger, D.; Hartmann, S.; Nordmeier, V.; Upmeyer zu Belzen, A. Measuring Scientific Reasoning Competencies: Multiple Aspects of Validity. In *Student Learning in German Higher Education*; Zlatkin-Troitschanskaia, O., Pant, H.A., Toepper, M., Lautenbach, C., Eds.; Springer Fachmedien Wiesbaden: Wiesbaden, Germany, 2020; pp. 261–280.
157. Krell, M.; Mathesius, S.; van Driel, J.; Vergara, C.; Krüger, D. Assessing Scientific Reasoning Competencies of Pre-Service Science Teachers: Translating a German Multiple-Choice Instrument into English and Spanish. *Int. J. Sci. Educ.* **2020**, *42*, 2819–2841. [CrossRef]
158. Mahler, D.; Bock, D.; Bruckermann, T. Preservice Biology Teachers Scientific Reasoning Skills and Beliefs about Nature of Science: How do They Develop and is There a Mutual Relationship during the Development? *Educ. Sci.* **2021**, *9*, 558. [CrossRef]
159. Kersting, N. Using Video Clips of Mathematics Classroom Instruction as Item Prompts to Measure Teachers' Knowledge of Teaching Mathematics. *Educ. Psychol. Meas.* **2008**, *68*, 845–861. [CrossRef]
160. Kunz, H.; Wolowski, J. Wissenschaftliches Denken und Arbeiten im Kompetenzorientierten Biologieunterricht-Aufbau von Fachmethodischem Wissen in der Qualifizierung Angehender Lehrkräfte. In *Vielfältige Wege biologiedidaktischer Forschung. Vom Lernort Natur zur Naturwissenschaftlichen Erkenntnisgewinnung in die Lehrerprofessionalisierung*; Meier, M., Wulff, C., Ziepprecht, K., Eds.; Waxmann: Münster, Germany, 2021; pp. 189–202.
161. Joachim, C.; Hammann, M.; Carstensen, C.H.; Bögeholz, S. Modeling and Measuring Pre-Service Teachers' Assessment Literacy Regarding Experimentation Competences in Biology. *Educ. Sci.* **2020**, *10*, 140. [CrossRef]
162. Van der Valk, T.; de Jong, O. Scaffolding Science Teachers in Open-Inquiry Teaching. *Int. J. Sci. Educ.* **2009**, *31*, 829–850. [CrossRef]
163. Hilfert-Rüppell, D.; Looß, M.; Klingenberg, K.; Eghtessad, A.; Höner, K.; Müller, R.; Strahl, A.; Pietzner, V. Scientific Reasoning of Prospective Science Teachers in Designing a Biological Experiment. *Lehrbild. Prüfstand* **2013**, *6*, 135–154.
164. Rauin, U.; Meier, U. Subjektive Einschätzungen des Kompetenzerwerbs in der Lehramtsausbildung. In *Forschung zur Lehrerbildung*; Lüders, M., Ed.; Waxmann: Münster, Germany, 2007; pp. 102–131.
165. Pugh, K.J.; Bergin, D.A. Motivational Influences on Transfer. *Educ. Psychol.* **2006**, *41*, 147–160. [CrossRef]
166. Simosi, M. The Moderating Role of Self-Efficacy in the Organizational Culture–Training Transfer Relationship. *Int. J. Train. Dev.* **2012**, *16*, 92–106. [CrossRef]
167. Schütze, B.; Rakoczy, K.; Klieme, E.; Besser, M.; Leiss, D. Training Effects on Teachers' Feedback Practice. The Mediating Function of Feedback Knowledge and the Moderating Role of Self-Efficacy. *ZDM* **2017**, *49*, 475–489. [CrossRef]
168. Fauth, B.; Decristan, J.; Decker, A.-T.; Büttner, G.; Hardy, I.; Klieme, E.; Kunter, M. The Effects of Teacher Competence on Student Outcomes in Elementary Science Education: The Mediating Role of Teaching Quality. *Teach. Teach. Educ.* **2019**, *86*, 102882. [CrossRef]
169. Khan, S.; Krell, M. Scientific Reasoning Competencies: A Case of Preservice Teacher Education. *Can. J. Sci. Math. Techn. Educ.* **2019**, *19*, 446–464. [CrossRef]

170. Vorholzer, A.; von Aufschnaiter, C.; Kirschner, S. Entwicklung und Erprobung eines Tests zur Erfassung des Verständnisses Experimenteller Denk- und Arbeitsweisen. *ZfDN* **2016**, *22*, 25–41. [CrossRef]
171. Lawson, A.E.; Clark, B.; Cramer-Meldrum, E.; Falconer, K.A.; Sequist, J.M.; Kwon, Y.-J. Development of Scientific Reasoning in College Biology: Do Two Levels of General Hypothesis-Testing Skills Exist? *J. Res. Sci. Teach.* **2000**, *37*, 81–101. [CrossRef]
172. Bosse, S.; Spörer, N. Erfassung der Einstellung und der Selbstwirksamkeit von Lehramtsstudierenden zum inklusiven Unterricht. *Empir. Sonderpädagogik* **2014**, *4*, 279–299.
173. Jerusalem, M.; Satow, L. Schulbezogene Selbstwirksamkeitserwartung. In *Skalen zur Erfassung von Lehrer- und Schülermerkmalen*; Jerusalem, M., Schwarzer, R., Eds.; Freie Universität Berlin: Berlin, Germany, 1999; p. 15.
174. Caspary, R.; Blanck, B.; Spychiger, M.; Bosch, B.; Steinbrink, D.; Beutelspacher, A.; Kahl, R.; Osten, M.; Schumacher, R.; Spitzer, M. Der Produktive Umgang mit Fehlern. In *Nur wer Fehler Macht, Kommt Weiter. Wege zu einer Neuen Lernkultur*; Herder: Freiburg, Germany, 2008; pp. 49–72.
175. Schlatter, E.; Lazonder, A.W.; Molenaar, I.; Janssen, N. Individual Differences in Children's Scientific Reasoning. *Educ. Sci.* **2021**, *11*, 471. [CrossRef]
176. Herppich, S.; Praetorius, A.-K.; Förster, N.; Glogger-Frey, I.; Karst, K.; Leutner, D.; Behrmann, L.; Böhmer, M.; Ufer, S.; Klug, J.; et al. Teachers' Assessment Competence: Integrating Knowledge-, Process-, and Product-Oriented Approaches into a Competence-Oriented Conceptual Model. *Teach. Teach. Educ.* **2018**, *76*, 181–193. [CrossRef]
177. Behrmann, L.; van Ophuysen, S. Das Vier-Komponenten-Modell der Diagnosequalität. In *Diagnostische Kompetenz von Lehrkräften: Theoretische und Methodische Weiterentwicklungen*; Pädagogische Psychologie und Entwicklungspsychologie; Südkamp, A., Praetorius, A.-K., Eds.; Waxmann: Münster, Germany; New York, NY, USA, 2017; pp. 38–41.
178. Vogt, F.; Schmiemann, P. Assessing Biology Pre-Service Teachers' Professional Vision of Teaching Scientific Inquiry. *Educ. Sci.* **2020**, *10*, 332. [CrossRef]

Article

Patterns of Scientific Reasoning Skills among Pre-Service Science Teachers: A Latent Class Analysis

Samia Khan ^{1,*} and Moritz Krell ² ¹ Department of Curriculum and Pedagogy, University of British Columbia, Vancouver, BC V5K 0A1, Canada² Leibniz Institute for Science and Mathematics Education, University of Kiel, 24103 Kiel, Germany; krell@leibniz-ipn.de

* Correspondence: samia.khan@ubc.ca

Abstract: We investigated the scientific reasoning competencies of pre-service science teachers (PSTs) using a multiple-choice assessment. This assessment targeted seven reasoning skills commonly associated with scientific investigation and scientific modeling. The sample consisted of 112 PSTs enrolled in a secondary teacher education program. A latent class (LC) analysis was conducted to evaluate if there are subgroups with distinct patterns of reasoning skills. The analysis revealed two subgroups, where LC1 (73% of the PSTs) had a statistically higher probability of solving reasoning tasks than LC2. Specific patterns of reasoning emerged within each subgroup. Within LC1, tasks involving analyzing data and drawing conclusions were answered correctly more often than tasks involving formulating research questions and generating hypotheses. Related to modeling, tasks on testing models were solved more often than those requiring judgment on the purpose of models. This study illustrates the benefits of applying person-centered statistical analyses, such as LC analysis, to identify subgroups with distinct patterns of scientific reasoning skills in a larger sample. The findings also suggest that highlighting specific skills in teacher education, such as: formulating research questions, generating hypotheses, and judging the purposes of models, would better enhance the full complement of PSTs' scientific reasoning competencies.

Citation: Khan, S.; Krell, M. Patterns of Scientific Reasoning Skills among Pre-Service Science Teachers: A Latent Class Analysis. *Educ. Sci.* **2021**, *11*, 647. <https://doi.org/10.3390/educsci11100647>

Keywords: scientific reasoning; science teacher education; pre-service teachers; person-centered statistical analyses; latent class analysis

Academic Editor: Ian Hay

Received: 17 August 2021
Accepted: 13 October 2021
Published: 15 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Scientific reasoning has been a subject of study in the field of science education for some time [1]. Assessing this reasoning, however, remains a 21st century challenge for science educators today [2]. The present study is on the scientific reasoning of future science teachers themselves. We have assessed reasoning amongst this group because they will need to teach and demonstrate reasoning to their future students in science, and we can design activities in science teacher education that can enhance their competency in this field.

Scientific reasoning is a competency that encompasses the abilities needed for scientific problem-solving, as well as the capacity to reflect on problem-solving [3,4]. In the sciences, reasoning has been previously distinguished from other constructs such as problem-solving and critical thinking or scientific thinking alone. Descriptions of thinking, problem-solving, and reasoning are often conflated. For example, scientific reasoning has been suggested as being a kind of problem-solving; however, it has also been suggested that reasoning can be distinguished from problem-solving alone in that direct retrieval of a solution from memory is not possible with reasoning [5]. Ford [6] further reinforces that reasoning does not mean following a series of rules either but rather encompasses permanent evaluation and critique, as suggested by the reflective component of the above definition. Reasoning in the sciences requires cognitive processes that can contribute to, or allow for, inquiring and answering questions about the world and the nature of phenomena. These cognitive processes include

formulating and evaluating hypotheses, two of several processes regularly invoked in scientific domains [7,8].

The multiple cognitive processes that have been investigated in research on reasoning in science and science education have been variously described as formal logic, non-formal reasoning, creativity, model-based reasoning, abductive reasoning, analogical reasoning, and probabilistic reasoning [9–12]. These processes may or may not be used in the wider category of critical thinking [13]. Scholars have provided evidence that the ability to use these processes for reasoning is transferable across domains [14], while others such as Kind and Osborne [15] suggest that reasoning is highly variable by the content and the procedural and epistemic knowledge of the reasoner. Scholars have also shown that the ability to reason in science does not necessarily improve with age [16] but that it can be taught and enhanced in both the early years and at university levels [17–19].

Our focus in the present study is on the reasoning competencies of pre-service science teachers (PSTs) enrolled in a university teacher education program. Most studies on pre-service science teachers' scientific reasoning competencies adopt variable-centered approaches and report, for example, average scores for sample groups or populations. For example, one study [20] reported on a group of 66 Australian pre-service science teachers that they performed significantly better on tasks that required skills of 'planning investigations' compared to tasks related to skills of 'formulating research questions' and 'generating hypotheses'. Such insights are valuable but sometimes might be too rough-grained depending on the research questions, as different subgroups with distinct patterns of scientific reasoning skills exist within a sample. In order to identify such subgroups, person-centered analyses are necessary, that, statistically speaking, aim to "[R]educe the 'noise' in the data by splitting the total variability into 'between-group' variability and 'within-group' variability" [21] (p. 2). Hence, person-centered analyses, like latent class analyses (LCA), are finer-grained analyses in the sense that they are case-based and identify individuals with similar patterns of scientific reasoning skills (e.g., [22]). Person-centered analyses are also referred to as 'typological' approaches [23]. Such approaches can be specifically valuable for educators as they move beyond the 'average' and follow, methodologically, "[M]odern developmental theory, in which individuals are regarded as the organising unit of human development" [23] (p. 502). In the present study, we seek to establish whether subgroups of reasoners can be ascertained among PSTs using an LCA. The seven reasoning skills examined are: *formulating research questions*, *generating hypotheses*, *planning investigations*, *analyzing data and drawing conclusions*, *judging the purpose of models*, *testing models*, and *changing models*. While historical examination of scientific work has revealed that practices such as thought experiments, analogies, and imagistic simulation are important to scientists' development of new concepts [24], these seven skills under investigation were identified as key empirical areas of inquiry in science education [25–29] and likely having been taught in undergraduate science programs [3].

2. Materials and Methods

2.1. Sample

A full cohort of 56 PSTs from a university in North America participated in this study. Their mean age was 27 years ($SD = 6.34$; mode = 23). Data collection was done in their science teacher education secondary methods course within a Bachelor of Education after-degree program. To enroll in the secondary program, all students had at least one prior degree (usually 4 years of Science or more). The instrument described below (Section 2.2) was administered to the PSTs in their methods course at the beginning and at the end of the semester (pre–post-assessment). For the purpose of identifying groups with distinct patterns of scientific reasoning, we analyzed pre- and post-assessment data taken together of 56 PSTs. The total response sample for each item was thus $n = n_{\text{pre}} + n_{\text{post}}$ or $n = 112$. Only PSTs without missing responses have been included in the analysis, resulting in a sample of $n = 101$ for the statistical analysis. The number of PSTs by primary major were: Biology ($n = 30$), Chemistry ($n = 11$), Physics ($n = 8$), Biomedicine ($n = 1$), Earth Sciences

($n = 1$), Mathematics ($n = 1$), n/a ($n = 4$). Most of the PSTs' prior degrees were within the field of Biology ($n = 60$; e.g., general Biology, Applied Biology, or Evolutionary Biology), followed by Chemistry ($n = 25$) and Physics ($n = 6$).

2.2. Data Collection

An established multiple-choice instrument was administered to assess the PSTs' scientific reasoning competencies. The instrument was originally developed in the German language [27] and was later adapted into English, with thorough evaluations [30]. The instrument includes 21 multiple-choice items that were developed to assess seven reasoning skills of *formulating research questions, generating hypotheses, planning investigations, analyzing data and drawing conclusions, judging the purpose of models, testing models, and changing models*. Authentic scientific contexts were included in the items, which are mostly related to general science and Biology as well. As suggested in the organizing device that has been used for test development (see Table 1), these seven skills are related to two sub-competencies: conducting scientific investigations and using scientific models [31]. To correctly solve the multiple-choice items, PSTs have to apply their procedural and epistemic knowledge related to the respective skills [32–34]. Table 1 lists the two sub-competencies, their associated skills, and the specific knowledge necessary to correctly answer the items.

Table 1. Sub-competencies of scientific reasoning and associated skills with necessary procedural and epistemic knowledge, as described by Mathesius et al. [34].

Sub-Competencies	Skills	Necessary Knowledge PSTs Have to Know That . . .
Conducting scientific investigations	formulating questions	... scientific questions are related to phenomena, empirically testable, intersubjectively comprehensible, unambiguous, basically answerable and are internally and externally consistent.
	generating hypotheses	... hypotheses are empirically testable, intersubjectively comprehensible, clear, logically consistent and compatible with an underlying theory.
	planning investigations	... causal relationships between independent and dependent variables based on a previous hypothesis can be examined, whereby the independent variable is manipulated during experiments and control variables are considered. ... correlative relationships between independent and dependent variables based on a previous hypothesis can be examined with scientific observations.
	analyzing data and drawing conclusions	... data analysis allows an evidence-based interpretation and evaluation of the research question and hypothesis.
Using scientific models	judging the purpose of models	... models can be used for hypotheses generation.
	testing models	... models can be evaluated by testing model-based hypotheses.
	changing models	... models are changed if model-based hypotheses are falsified.

2.3. Data Analysis: Latent Class Analysis

A latent class analysis (LCA) was utilized to identify patterns of scientific reasoning skills among PSTs. The R package *poLCA* was employed [35]. All further (classical) statistical analyses, such as *t*-tests and descriptive analyses, were carried out with IBM

SPSS statistics, version 26. In an LCA, PSTs' responses are analyzed on the latent level, all variables are assumed to be (at least) on a nominal level, and there are no restrictions on the kind of relation between the (manifest) variables [33,36,37]. LCA was selected for data analysis because it permits the identification and computation of different groups (i.e., latent classes) of PSTs, with each group consisting of individuals with a response pattern that is as homogenous as possible (low within-group variability) but different from the response patterns of the other groups (high between-group variability). Therefore, LCA would be considered as belonging to the person-centered approaches of data analyses [21,23].

A core question of LCA is to decide on the appropriate number of latent classes [36]. To compare different LCA models, indices such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), and the sample size adjusted Bayesian information criterion (ssaBIC) are typically employed. These indices factor in the parsimony, the sample size, and the likelihood of the LCA models—each of the indices in a different manner [38]. When comparing different LCA models with these information indices, the smallest value of each index points out the comparatively best LCA model; however, the BIC and the ssaBIC were identified as superior indicators compared to the AIC [39] (p. 557), which is why these indicators are used in the present study. On the other hand, the BIC and the ssaBIC often do not identify the same LCA model as optimal [38]. Therefore, one has to use a combination of different insights to decide how many latent classes represent the data set best [38].

It is an important characteristic of LCA that the subjects are not assigned to the different latent classes in a deterministic manner but more so in a probabilistic sense. For diagnostic purposes, it is common to classify each subject to the latent class with the highest probability of assignment. Therefore, an “Additional indicator [of model-goodness] is the average membership probability within each [latent] class” [40] (p. 52); the higher this probability, the better the LCA model. Furthermore, one should analyze the item parameters for extreme values that indicate an estimated probability of 0% or 100% to solve a task; the fewer extreme values, the better the LCA model [40].

3. Results

Table 2 provides the fit-indices for the LCA models compared in this study. Because the BIC (2 latent classes) and ssaBIC (4 latent classes) suggest selecting different LCA models, the number of extreme values and the probability of assignment have been used as additional indicators. Based on these indicators, it can be assumed that the response pattern of the PSTs is best represented using two latent classes. These two latent classes consist of about 73% or 74 PSTs (latent class 1) and 27% or 27 PSTs (latent class 2) of the sample, respectively.

Table 2. Fit-indices of the different LCA models compared. Note that models with more than four latent classes did not fit the data.

LCA Model	BIC	ssaBIC	Extreme Values	Probability of Assignment
2 latent classes	2685	2549	0	0.93 to 0.98
3 latent classes	2722	2517	9	0.92 to 0.98
4 latent classes	2779	2504	11	0.91 to 0.97

Figure 1 illustrates the response profiles for the two latent classes across the seven skills of scientific reasoning covered in the multiple-choice instrument. Generally, PSTs in latent class 1 show a higher mean probability of correct answers within all seven skills. Comparing the mean probability of correct answers between the two latent classes with independent *t*-tests resulted in significant differences for the skills *planning investigations* ($p = 0.04$; $d = 0.48$, small to medium effect size measure), *analyzing data and drawing conclusions* ($p < 0.001$; $d = 1.25$, large effect size measure) as well as *judging the purpose of models*

($p < 0.001$; $d = 1.25$, large effect size measure), *testing models* ($p < 0.001$; $d = 1.49$, large effect size measure), and *changing models* ($p < 0.001$; $d = 0.88$, large effect size measure).

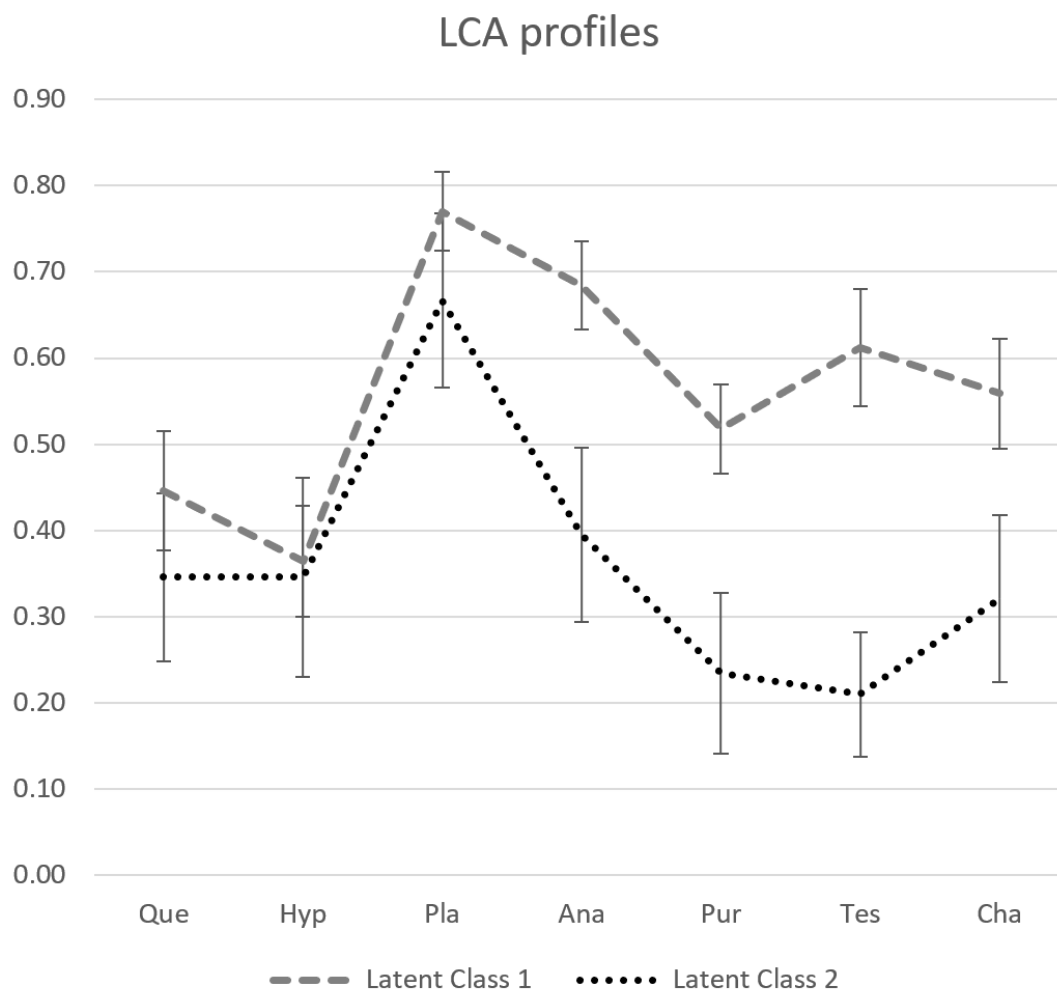


Figure 1. Response profiles for the two latent classes across the seven skills of scientific reasoning (mean score ± 2 * standard error).

For latent class 1 and considering skills related to conducting scientific investigations (Table 1), response probabilities for the skills *formulating research questions* and *generating hypotheses* on the one hand, and *planning investigations* and *analyzing data and drawing conclusions*, on the other hand, are quite similar, even though significant differences with large effect size measures could be found between these two groups of skills. For the skills related to using scientific models (Table 1), correct responses were found significantly more often for the skill *testing models* than for *judging the purpose of models* ($p = 0.02$; $d = 0.36$, small effect size measure).

For latent class 2 and considering skills related to conducting scientific investigations (Table 1), items related to the skill *planning investigations* have been answered correctly significantly more often than the tasks related to the other three skills ($p < 0.001$; $d > 1.00$, large effect size measures). For using scientific models (Table 1), no significant differences between the skills could be found.

In order to better understand the characteristics of the PSTs assigned to latent class 1 and latent class 2, we compared their age, primary majors, and the sum of previous degrees. Independent *t*-tests (Table 3) revealed that there are significantly more PSTs with the primary major of Biology in latent class 1 (about 65%) than in latent class 2 (about 33%). For the primary major of Chemistry, it is quite the reverse (about 15% in latent class 1 and about 33% in latent class 2); also, the number of PSTs with more than one previous degree

is significantly higher in latent class 1 ($n = 11$) than in latent class 2 ($n = 1$). These findings illustrate that the study of Biology as a primary major and a higher number of previous degrees made it more likely to belong to the more proficient latent class 1, whereas the study of Chemistry as a primary major made it more likely to belong to latent class 2.

Table 3. Comparison of the PSTs assigned to latent class (LC) 1 and LC 2 along the variables age, primary major of Biology, Chemistry or Physics, and the sum of previous degrees (the latter as a dichotomized variable with 1 = one previous degree and 2 = more than one previous degree).

Variable	LC Assignment	N	M	SD	t-Test
Age	1	74	26.54	5.35	$t(99) = 0.591; p = 0.556$
	2	27	27.30	6.55	
Biology	1	74	0.65	0.48	$t(99) = 2.918; p = 0.004$
	2	27	0.33	0.48	
Chemistry	1	74	0.15	0.36	$t(37.07) = 1.821; p = 0.077 *$
	2	27	0.33	0.48	
Physics	1	74	0.14	0.34	$t(99) = 0.316; p = 0.753$
	2	27	0.11	0.32	
Previous degrees	1	66	1.24	0.63	$t(78.93) = 2.072; p = 0.042 *$
	2	25	1.08	0.40	

* Adjusted t -statistic and df because of violated assumption of variance homogeneity.

4. Discussion

Using LCA, we revealed that two groups of reasoners emerged amongst the PSTs. One subgroup (latent class 1) had a statistically higher probability of solving scientific reasoning tasks than the other subgroup (latent class 2). Overall, the groups were significantly different on the following five skills out of seven investigated: *planning investigations*, *analyzing data and drawing conclusions*, *judging the purpose of models*, *testing models*, and *changing models*. They were not significantly different from each other on *formulating research questions* and *generating hypotheses*.

The latent class 1 subgroup responded significantly differently from each other on the skills *planning investigations* and *analyzing data and drawing conclusions* in contrast to the skills *formulating research questions* and *generating hypotheses*. Tasks about *testing models* were solved more often than those requiring *judging the purpose of models* within this subgroup. The latent class 2 subgroup responded significantly differently from each other on *planning investigations* compared to the other skills. For using scientific models, no significant differences could be found within this subgroup on the skills related to modeling (*judging the purpose of models*, *testing models*, and *changing models*).

These two subgroups also shared several other key characteristics. In latent class 1, a significant majority had a major in Biology compared to latent class 2, whereas there were far fewer from Chemistry in latent class 1. Moreover, there were significantly more PSTs with more than one previous degree in latent class 1 than in latent class 2. This finding is noteworthy for science teacher education because it suggests that Biology majors were significantly better at *planning investigations*, *analyzing data and drawing conclusions*, *judging the purpose of models*, *testing models*, and *changing models* than Chemistry majors. These findings might have been caused by the dominance of Biology-related items in the instrument; however, as the items require PSTs to apply procedural and epistemic knowledge as shown in Table 1 (and less so content knowledge), the findings lead us towards a renewed emphasis on reasoning tasks for Chemistry teacher education. Nevertheless, future studies could investigate the importance of science content knowledge from specific subjects (such as Biology) for solving the items, for instance, by applying think-aloud studies [25] or statistically investigating difficulty-generating task characteristics [41].

As a 'person-centered' statistical approach, the LCA was particularly powerful in ascertaining subgroups within a science teacher education cohort. This statistical approach

is a departure from traditional variable-centered approaches in education that tend to report on average scores for sample groups [21,23]. The LCA permits statistical cases to emerge from within samples or classrooms and is a recommended approach to generate case studies for further inquiry in science teacher education research.

In combination with relevant epistemic, procedural, and content knowledge, greater attention to *formulating research questions* and *generating hypotheses* would be helpful within science teacher education. Furthermore, reasoning tasks involving *judging the purpose of models* and *changing models* could be a high priority for modeling investigations in pre-service science teacher education. Possible science teacher education activities to support such tasks include the three-phased generating, evaluating, and modifying (GEM) models approach [10]. This approach emphasizes generating hypotheses in the first phase and testing and changing models in the second and third phases [42]. In general, science teacher education courses, Biology majors, or those with additional degrees could be purposefully included within heterogeneous groups for cooperative learning tasks. It was interesting to the authors that Biology majors outperformed other majors in this study, although this might be caused by the dominance of Biology-related items in the instrument; insights into the differences in performance among majors would be a helpful avenue for the design of science teachers education courses and group work in the ways suggested above. By participating in reasoning tasks with such recommendations in mind, future teachers might be able to better support their own students to develop competencies in these areas.

The significance of this study is that it identifies two groups of reasoners who are PSTs with different propensities to reason in science using person-centered statistics. Normally, the classroom would be treated similarly as an entire group; however, with this statistical approach, the researchers are able to show that subgroups of PSTs themselves emerged as competent at very different reasoning tasks. One subgroup is significantly more competent at *planning investigations, analyzing data and drawing conclusions, judging the purpose of models, testing models, and changing models* than the other. The subgroups had approximately equivalent competencies at *formulating research questions* and *generating hypotheses* showing for the first time that among PSTs, different subgroups with specific patterns of scientific reasoning skills exist. This finding can have an impact on science students of these future teachers, who presumably will draw upon their own competencies to demonstrate how to reason in the classroom. Future directions for research could target investigation and model-based reasoning competencies among PSTs and relationships to student reasoning. *Judging the purpose of models, formulating research questions, and generating hypotheses* were areas that PSTs were less competent; researching interventions related to these aspects of modeling and investigation would be worthwhile.

Author Contributions: Conceptualization, M.K., S.K.; methodology, M.K.; investigation, M.K., S.K.; resources, M.K., S.K.; writing—original draft preparation, S.K.; writing—review and editing, M.K., S.K.; visualization, M.K.; funding acquisition, M.K., S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the 2018 UBC-FUB Joint Funding Scheme, grant number FSP-2018-401.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of the University of British Columbia (ID H18-01801, approved 23 July 2018).

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

Data Availability Statement: Data is available upon request to the second author.

Acknowledgments: The authors wish to thank Alexis Gonzalez for support in data collection and tabulation.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Lawson, A.E. The development and validation of a classroom test of formal reasoning. *J. Res. Sci. Teach.* **1978**, *15*, 11–24. [CrossRef]
2. Osborne, J. The 21st century challenge for science education: Assessing scientific reasoning. *Think. Ski. Creat.* **2013**, *10*, 265–279. [CrossRef]
3. Khan, S.; Krell, M. Scientific reasoning competencies: A case of preservice teacher education. *Can. J. Sci. Math. Technol. Educ.* **2019**, *19*, 446–464. [CrossRef]
4. Lawson, A.E. The nature and development of scientific reasoning: A synthetic view. *Int. J. Sci. Math. Educ.* **2004**, *2*, 307–338. [CrossRef]
5. Zimmerman, C. The development of scientific reasoning skills. *Dev. Rev.* **2000**, *20*, 99–149. [CrossRef]
6. Ford, M.J. Educational implications of choosing “practice” to describe science in the next generation science standards. *Sci. Educ.* **2015**, *99*, 1041–1048. [CrossRef]
7. Díaz, C.; Dorner, B.; Hussmann, H.; Strijbos, J.W. Conceptual review on scientific reasoning and scientific thinking. *Curr. Psychol.* **2021**, 1–13.
8. Reith, M.; Nehring, A. Scientific reasoning and views on the nature of scientific inquiry: Testing a new framework to understand and model epistemic cognition in science. *Int. J. Sci. Educ.* **2020**, *42*, 2716–2741. [CrossRef]
9. Krell, M.; Hergert, S. Modeling strategies. In *Towards a Competence-Based View on Models and Modeling in Science Education*; Upmeier zu Belzen, A., Krüger, D., van Driel, J., Eds.; Springer: Cham, Switzerland, 2020; pp. 147–160.
10. Khan, S. Model-based inquiries in chemistry. *Sci. Educ.* **2007**, *91*, 877–905. [CrossRef]
11. Babai, R.; Brecher, T.; Stavvy, R.; Tirosh, D. Intuitive interference in probabilistic reasoning. *Int. J. Sci. Math. Educ.* **2006**, *4*, 627–639. [CrossRef]
12. Upmeier zu Belzen, A.; Engelschalt, P.; Krüger, D. Modeling as scientific reasoning—The role of abductive reasoning for Modeling competence. *Educ. Sci.* **2021**, *11*, 495. [CrossRef]
13. Holyoak, K.J.; Morrison, R.G. Thinking and reasoning: A reader’s guide. In *Oxford Handbook of Thinking and Reasoning*; Holyoak, K.J., Morrison, R.G., Eds.; Oxford University Press: New York, NY, USA, 2005.
14. Kuhn, D.; Arvidsson, T.S.; Lesperance, R.; Corprew, R. Can engaging in science practices promote deep understanding of them? *Sci. Educ.* **2017**, *101*, 232–250. [CrossRef]
15. Kind, P.; Osborne, J. Styles of scientific reasoning: A cultural rationale for science education? *Sci. Educ.* **2017**, *101*, 8–31. [CrossRef]
16. Kuhn, D. What is scientific thinking and how does it develop? In *Handbook of Childhood Cognitive Development*; Goswami, U., Ed.; Blackwell: Oxford, UK, 2002; pp. 371–393.
17. Schauble, L. In the eye of the beholder: Domain-general and domain-specific reasoning in science. In *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*; Fischer, F., Chinn, C., Engelmann, K., Osborne, J., Eds.; Routledge: New York, NY, USA, 2018.
18. Dunbar, K.; Klahr, D. *Developmental Differences in Scientific Discovery Processes*; Psychology Press: Hove, UK, 2013; pp. 129–164.
19. Morris, B.J.; Croker, S.; Masnick, A.M.; Zimmerman, C. The emergence of scientific reasoning. In *Current Topics in Children’s Learning and Cognition*; Kloos, H., Morris, B., Amaral, J., Eds.; IntechOpen: London, UK, 2012; pp. 61–82.
20. Krell, M.; Dawborn-Gundlach, M.; van Driel, J. Scientific reasoning competencies in science teaching. *Teach. Sci.* **2020**, *66*, 32–42.
21. Kusurkar, R.A.; Mak-van der Vossen, M.; Kors, J.; Grijpma, J.W.; van der Burgt, S.M.; Koster, A.S.; de la Croix, A. ‘One size does not fit all’: The value of person-centred analysis in health professions education research. *Perspect. Med. Educ.* **2020**, *10*, 245–251. [CrossRef] [PubMed]
22. Krell, M.; zu Belzen, A.U.; Krüger, D. Students’ levels of understanding models and modeling in biology: Global or aspect-dependent? *Res. Sci. Educ.* **2014**, *44*, 109–132. [CrossRef]
23. Watt, H.M.; Parker, P.D. Person-and variable-centred quantitative analyses in educational research: Insights concerning Australian students’ and teachers’ engagement and wellbeing. *Aust. Educ. Res.* **2020**, *47*, 501–515. [CrossRef]
24. Nersessian, N.J. *Creating Scientific Concepts*; MIT Press: Cambridge, MA, USA, 2010.
25. Krell, M.; Redman, C.; Mathesius, S.; Krüger, D.; van Driel, J. Assessing pre-service science teachers’ scientific reasoning competencies. *Res. Sci. Educ.* **2020**, *50*, 2305–2329. [CrossRef]
26. Krell, M. Schwierigkeitserzeugende Aufgabenmerkmale bei Multiple-Choice-Aufgaben zur Experimentierkompetenz im Biologieunterricht: Eine Replikationsstudie [Difficulty-creating task characteristics in multiple-choice questions on experimental competence in biology classes: A replication study]. *Z. Didakt. Nat.* **2018**, *24*, 1–15.
27. Krüger, D.; Hartmann, S.; Nordmeier, V.; Upmeier zu Belzen, A. Measuring scientific reasoning competencies. In *Student Learning in German Higher Education*; Springer: Wiesbaden, Germany, 2020; pp. 261–280. [CrossRef]
28. Opitz, A.; Heene, M.; Fischer, F. Measuring scientific reasoning—A review of test instruments. *Educ. Res. Eval.* **2017**, *23*, 78–101. [CrossRef]
29. Bicak, B.E.; Borchert, C.E.; Höner, K. Measuring and Fostering Preservice Chemistry Teachers’ Scientific Reasoning Competency. *Educ. Sci.* **2021**, *11*, 496. [CrossRef]

30. Krell, M.; Mathesius, S.; van Driel, J.; Vergara, C.; Krüger, D. Assessing scientific reasoning competencies of pre-service science teachers: Translating a German multiple-choice instrument into English and Spanish. *Int. J. Sci. Educ.* **2020**, *42*, 2819–2841. [CrossRef]
31. Hartmann, S.; Upmeier zu Belzen, A.; Krüger, D.; Pant, H.A. Scientific reasoning in higher education. *Z. Psychol.* **2015**, *223*, 47–53. [CrossRef]
32. Mathesius, S.; Krell, M. Assessing modeling competence with questionnaires. In *Towards a Competence-Based View on Models and Modeling in Science Education*; Upmeier zu Belzen, A., Krüger, D., van Driel, J., Eds.; Springer: Cham, Switzerland, 2020; pp. 117–129.
33. Hagenaars, J.; Halman, L. Searching for ideal types: The potentialities of latent class analysis. *Eur. Sociol. Rev.* **1989**, *5*, 81–96. [CrossRef]
34. Mathesius, S.; Upmeier zu Belzen, A.; Krüger, D. Competencies of biology students in the field of scientific inquiry: Development of a testing instrument. *Erkenn. Biol.* **2014**, *13*, 73–88.
35. Linzer, D.; Lewis, J. poLCA: Polytomous Variable Latent Class Analysis. R Package Version 1.4. 2013. Available online: <http://dlinzer.github.com/poLCA> (accessed on 7 December 2020).
36. Collins, L.; Lanza, S. *Latent Class and Latent Transition Analysis*; Wiley: Hoboken, NJ, USA, 2010.
37. Langeheine, R.; Rost, J. (Eds.) *Latent Trait and Latent Class Models*; Plenum Press: New York, NY, USA, 1988.
38. Henson, J.; Reise, S.; Kim, K. Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Struct. Equ. Model.* **2007**, *14*, 202–226. [CrossRef]
39. Nylund, K.; Asparouhov, T.; Muthén, B. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Struct. Equ. Model.* **2007**, *14*, 535–569. [CrossRef]
40. Spiel, C.; Glück, J. A model-based test of competence profile and competence level in deductive reasoning. In *Assessment of Competencies in Educational Contexts*; Hartig, J., Klieme, E., Leutner, D., Eds.; Hogrefe & Huber: Göttingen, Germany, 2008; pp. 45–68.
41. Krell, M.; Khan, S.; van Driel, J. Analyzing Cognitive Demands of a Scientific Reasoning Test Using the Linear Logistic Test Model (LLTM). *Educ. Sci.* **2021**, *11*, 472. [CrossRef]
42. Khan, S. New pedagogies on teaching science with computer simulations. *J. Sci. Educ. Technol.* **2011**, *20*, 215–232. [CrossRef]

Article

Analyzing Cognitive Demands of a Scientific Reasoning Test Using the Linear Logistic Test Model (LLTM)

Moritz Krell ^{1,*} , Samia Khan ² and Jan van Driel ³¹ IPN-Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, D-24118 Kiel, Germany² Department of Curriculum and Pedagogy, Faculty of Education, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; samia.khan@ubc.ca³ Melbourne Graduate School of Education, The University of Melbourne, Melbourne, VIC 3010, Australia; j.vandriel@unimelb.edu.au

* Correspondence: krell@leibniz-ipn.de

Abstract: The development and evaluation of valid assessments of scientific reasoning are an integral part of research in science education. In the present study, we used the linear logistic test model (LLTM) to analyze how item features related to text complexity and the presence of visual representations influence the overall item difficulty of an established, multiple-choice, scientific reasoning competencies assessment instrument. This study used data from $n = 243$ pre-service science teachers from Australia, Canada, and the UK. The findings revealed that text complexity and the presence of visual representations increased item difficulty and, in total, contributed to 32% of the variance in item difficulty. These findings suggest that the multiple-choice items contain the following cognitive demands: encoding, processing, and combining of textually presented information from different parts of the items and encoding, processing, and combining information that is presented in both the text and images. The present study adds to our knowledge of which cognitive demands are imposed upon by multiple-choice assessment instruments and whether these demands are relevant for the construct under investigation—in this case, scientific reasoning competencies. The findings are discussed and related to the relevant science education literature.

Keywords: scientific reasoning; cognition; assessment; item features; item difficulty

Citation: Krell, M.; Khan, S.; van Driel, J. Analyzing Cognitive Demands of a Scientific Reasoning Test Using the Linear Logistic Test Model (LLTM). *Educ. Sci.* **2021**, *11*, 472. <https://doi.org/10.3390/educsci11090472>

Academic Editor: Silvija Markic

Received: 15 July 2021

Accepted: 23 August 2021

Published: 27 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

An understanding of science and its procedures, capabilities, and limitations is crucial for a society facing complex problems. This significance was recently highlighted during the COVID-19 crisis, where misinformation through traditional and social forms of media appeared to be highly influential in shaping peoples' opinions and actions about the crisis [1]. Science education can respond to these issues in part by supporting the development of scientific reasoning competencies (SRC) among students of science. Additionally, science teachers would benefit from strong SRC themselves to model and promote SRC among their students [2–4]. SRC are defined as the dispositions to be able to solve a scientific problem in a certain situation by applying a set of scientific skills and knowledge, and by reflecting on the process of scientific problem-solving at a meta-level [5–8]. SRC are also seen as a core element of 21st-century skills in science curricula, as they are assumed to help enable civic participation in socio-scientific issues facing societies and have been said to be indicative of a society's future economic power [9,10]. Hence, SRC, such as developing scientific questions and hypotheses, modeling, generating evidence through experimentation, and evaluating claims, are addressed in science education policy papers and curriculum documents as a key outcome of science education in various countries around the world (e.g., [11–13]). SRC are also emphasized as part of science teachers' professional competencies that should be developed during initial teacher education [14].

Existing studies suggest that pre-service science teachers typically have basic SRC, with pre-service secondary teachers outperforming pre-service primary or early childhood teachers [5]. For the specific skill of scientific modeling, it was shown that pre-service science teachers apply strategies and experience challenges similar to secondary school students [15]. Furthermore, longitudinal studies revealed that SRC slightly develop during science teacher education at university [16] and that specific teacher education programs can contribute to competence development in this field [17].

The development and evaluation of assessments that are capable of providing valid measures of respondents' SRC have become an integral part of research in science education [8,18]; however, several authors have recently questioned the quality of many existing instruments to assess SRC. For example, Ding et al. [19] identified poor definitions of the underlying constructs to be measured and criticized that most scientific reasoning instruments, "[A]re in fact intended to target a broader construct of scientific literacy" (p. 623) rather than specific competencies needed for reasoning in science. In a review study, it was found that the psychometric quality of most published instruments to assess SRC was not evaluated satisfactorily [18]. Furthermore, Osborne [8] criticized a general lack of validity evidence for these available instruments and referred to the valid assessment of SRC, as, "[T]he 21st century challenge for science education."

Arguably, an exception to these criticisms regarding the quality of instruments to assess SRC is a German multiple-choice instrument that has recently been developed to assess pre-service science teachers' SRC during their course of studies at university [16,20]. English and Spanish adaptations of this instrument have also been developed and evaluated [5,21]. For the original German instrument, comprehensive sources of validity evidence have been considered following the recommendations in the Standards for Educational and Psychological Testing [22]. For example, the instrument has been developed based on a clear theoretical framework, distinguishing between two sub-competencies of scientific reasoning—*conducting scientific investigations* and *using scientific models*—and seven related skills of *formulating research questions, generating hypotheses, planning investigations, analyzing data and drawing conclusions, judging the purpose of models, testing models, and changing models*. Furthermore, standardized construction guidelines for item development were used based on this framework [23], and the whole process of item development was guided by a critical examination of various sources of validity evidence (e.g., [23,24]), as summarized in [16]. In this process, one validation study [24] analyzed the influence of item features on item difficulty. The authors found that item length (word count) and the use of visual images, tables, formulas, abstract concepts, and specialized terms in the items significantly contributed to item difficulty. Taken together, these features contributed to 32% of the variance in item difficulty. The authors argued that these findings still provide evidence for the valid interpretation of the test scores as measures of SRC because the identified effects of item features on item difficulty were in accordance with the theoretical background of item development, and they showed a plausible pattern of cognitive demands [24].

In general, the analysis of item features and their influence on item difficulty is a common approach to research in psychological and educational assessment [25–28]. The basic assumption in this context is that assessments should represent the construct under investigation and test items should stimulate cognitive processes that constitute the target construct (construct validity or construct representation, respectively, [29,30]). For example, items that are intended to assess the competencies of "analyzing evidence" might provide an experimental design and corresponding findings and ask students to interpret the evidence appropriately [28]. The development of test items has to account for item features and underlying cognitive processes so that the instrument allows for valid interpretations of obtained test scores [27]. Related to this, legitimate and illegitimate sources of item difficulty have been distinguished [24]. While legitimate sources of item difficulty are those that are intentionally implemented to assess skills or knowledge representative of the respective competency, illegitimate sources of item difficulty are not directly related to the target construct, such as reading capabilities in science or mathematics tests, and can negatively

impact valid test score interpretation [24]. Identifying threats to validity, such as construct-irrelevant sources of item difficulty, however, has the potential to inform item development and thus improve the validity of assessments. Furthermore, construct-relevant sources of item difficulty can guide item development [27,31]. Nonetheless, “[W]hat constitutes construct-irrelevant variance is a tricky and contentious issue” [30] (p. 743) and depends on the definition of the respective construct. As a result, exploratory studies investigating the influence of item features on item difficulty of an existing assessment instrument can contribute to a better understanding of the cognitive demands of the instrument [26,28].

This study adds to this body of research by investigating the influence of item features on item difficulty of the above-mentioned German multiple-choice instrument. This study contributes to construct validation of this internationally employed testing instrument [16,21]. Furthermore, and independent from the specific instrument, this study provides insights about the influence of item features on item difficulty, and as a result, might be used by scholars to provide direction for systematically developing testing instruments that account for such features [27]. The focus of this study is on formal item features related to text complexity and the presence of external visual representations. There are already some studies that investigated the influence of formal item features on item difficulty in science education. For example, text length has been identified as a feature that tends to increase item difficulty [24,32]. In contrast to internal (i.e., mental) representations, external representations are defined as externalizations or materializations of more or less abstract thoughts in the form of gestures, objects, pictures and signs [33]. Taxonomies of (external) representations distinguish between descriptions and depictions, with descriptions including text, mathematical expressions, and formulas and depictions including photographs, maps, and diagrams [34]. Many representations are also combinations of different forms. For example, diagrams include textual (descriptive) and graphical (depictive) elements [35]. Formal item features, such as text length or task format, have been described as being part of the surface structure of test items; that is, such item features are often not directly related to the construct to be assessed [32,36]. On the other hand, the existence of formal item features is an inevitable part of item development, and hence, knowledge about how such features influence item difficulty is of significance for scholars interested in developing testing instruments.

2. Aims of the Study and Hypotheses

This study investigates the effect of item features on item difficulty for the English adaptation of the multiple-choice SRC assessment instrument described above. Item features related to text complexity and the presence of visual representations will be tested for their influence on item difficulty. This study complements existing evaluation studies on the English adaptation of the instrument that have not yet analyzed item features [5,21]. Furthermore, the present study also significantly adds to our knowledge of which cognitive demands appear to be imposed upon by multiple-choice assessment instruments and whether these demands are relevant for the construct under investigation—in this case, SRC [24,28,31].

The following assumptions undergird the study: (1) item difficulty is increased with an increase in the complexity of text included in the item because the complex text makes it more difficult to encode and process information relevant to identify the attractor (or the correct answer option) [24,32]; (2) item difficulty is increased for items that contain visual representations next to textual information because this addition requires respondents to simultaneously encode and process information that is presented in text and image, which, in turn, increases cognitive load [37].

3. Materials and Methods

3.1. Sample and Data Collection

Data of $N = 243$ pre-service science teachers from Australia ($n = 103$; mean age = 28), Canada ($n = 112$; mean age = 27), and the UK ($n = 26$; mean age = 31) were analyzed in this

study. Some data partly originate from existing studies [2,3,5,21] and were secondarily analyzed for the purpose of this study. The UK sub-sample contains new data that have neither been analyzed nor published. Hence, this study made use of some available data sets in order to test the above hypotheses. Having an international sample with participants from three countries allowed the hypotheses to be tested independently from the specific context and, thus, potentially provide more generalizable findings. SRC are an important goal of science teacher education in all three countries [2,3].

In each case, participating pre-service science teachers voluntarily agreed to participate in this study and anonymously completed the instrument, which is why the sample sizes are relatively small (e.g., $n = 26$ from the UK). The study information was shared with participants digitally (i.e., via email) or in person, in science methods courses of the respective pre-service teacher education programs. Completing the instrument, however, occurred outside of courses and was not an obligatory part of the pre-service science teachers' curriculum. Ethics approval was also obtained from local ethics approval committees. To ensure equivalence of testing conditions, the same standardized test instruction was used in all three subsamples—namely, background information about the study and the assessed competencies, and voluntary participation.

In all three subsamples, the above-mentioned English adaptation of the German SRC assessment instrument was administered. As described in [5,21], the English adaptation was systematically translated and evaluated based on the German original instrument [16]. For each of the seven skills of *formulating research questions, generating hypotheses, planning investigations, analyzing data and drawing conclusions, judging the purpose of models, testing models, and changing models*, the English instrument includes three multiple-choice items (i.e., 21 items in total). Each item is contextualized within an authentic scientific context, and the respondents have to apply their procedural and epistemic knowledge within this context to identify the attractor. (For sample items, see [21]; the full instrument is available upon request to the first author).

3.2. Item Analysis

The aim of this study was to analyze the influence of item features related to text complexity and the presence of visual representations on item difficulty. For this purpose, 21 items were analyzed by a trained student assistant and the first author to obtain information about text complexity and the presence of visual representations (i.e., figures or diagrams) in each item. The latter was scored with yes (=1) or no (=0) as this scoring was also conducted in earlier studies (e.g., [24,32]). For text complexity, three different readability measures were calculated, as described in [38]: the 4. Wiener Sachtextformel (WSTF), local substantival textual cohesion (LSTC), and global substantival textual cohesion (GSTC). These readability measures provide a sound statistical estimation of text complexity in science education [38].

The 4. Wiener Sachtextformel (WSTF) calculates a readability measure based on the percentage of words with more than two syllables (SYLL) and the average length (i.e., word count) of sentences (SENT) as follows [39]:

$$\text{WSTF} = 0.2656 \cdot \text{SENT} + 0.2744 \cdot \text{SYLL} - 1.693. \quad (1)$$

Substantival textual cohesion indicates text coherence based on substantives, either locally (i.e., in consecutive sentences) or globally (i.e., in the whole text) [40]. Global substantival textual cohesion (GSTC) is calculated by dividing the number of substantives that appear more than once in a text (SUB_2) by the number of substantives that appear only once (SUB). Local substantival textual cohesion (LSTC) is calculated by dividing the number of substantially connected sentences (LSCS, i.e., consecutive sentences with the same substantive) by the total number of sentences (S) as follows:

$$\text{GSTC} = \frac{\text{SUB}_2}{\text{SUB}} \cdot 100\%, \quad (2)$$

$$\text{LSTC} = \frac{\text{LSCS}}{S} \cdot 100\%. \quad (3)$$

Higher numbers of WSTF and lower numbers of LSTC and GSTC indicate more complex texts; $5.4 < \text{WSTF} < 8.4$, $0.41 < \text{LSTC} < 0.65$, and $0.70 < \text{GSTC} < 0.89$ have been suggested as indicating appropriately understandable texts for science education [38].

3.3. Data Analysis: Linear Logistic Test Model

To estimate the influence of the different item features on an item's difficulty, the linear logistic test model (LLTM) was applied [41,42] as this model was applied in several similar studies analyzing item features (e.g., [28,43]). The LLTM belongs to Rasch models, a family of established psychometric models utilized in psychological and educational research [44]. The family of Rasch models includes descriptive and explanatory psychometric models [45,46]. For example, the one-parameter logistic model (1PLM) is a descriptive psychometric model that allows for the estimation of individual person ability (θ_s) and item difficulty (β_i) parameters. In 1PLM, it is assumed that the probability of a correct item response depends only on θ_s and β_i [44].

$$P(X_{is}) = \frac{\exp(\theta_s - \beta_i)}{1 + \exp(\theta_s - \beta_i)} \quad (4)$$

In contrast to descriptive models such as 1PLM, explanatory models consider item or person features to further explain the item difficulty or person ability parameters, respectively [46]. The LLTM is an item explanatory model because it assumes that item difficulty is a linear (additive) combination of basic parameters α_k [43]. Formally, the β_i parameter of 1PLM is replaced with a linear combination of these basic parameters [41] as follows:

$$\beta'_i = \sum_{k=1}^N (\alpha_k \chi_{ik}) \quad (5)$$

where α_k as the regression coefficient for k (i.e., the estimated difficulty of the item feature k), and χ_{ik} as the given weight of item feature k on item i (i.e., the extent to which the respective item feature applies to item i). Hence, α_k illustrates the contribution of item feature k to item difficulty [43]. If an LLTM can be shown to fit the given data, the estimated parameters α_k provide measures for the item features' contribution to item difficulty. More specifically, it is assumed that item difficulty can be sufficiently and totally explained with the specified parameters in the LLTM [42]. Therefore, the LLTM can be considered more restrictive and more parsimonious than the 1PLM [47].

To evaluate the model fit of an LLTM, a two-step procedure is proposed: first, 1PLM has to fit "at least approximately" [42] (p. 509) to the data. For testing the fit of a Rasch model to the given data, fit indices such as the sum of squared standardized residuals (MNSQs) are proposed. MNSQs provide a measure of the discrepancy between the assumptions of the Rasch model and the observed data [48]. Second, the decomposition of β_i (Formula 5) needs to be checked for empirical validity. For this reason, the item difficulty parameters estimated in 1PLM, and the corresponding LLTM can be compared (e.g., graphically or by calculating Pearson correlation coefficient, [25]). High associations between both parameters indicate that the decomposition of β_i might be valid [42]. Furthermore, information criteria, such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), and the log-likelihood difference test can be applied to compare the fit of both models and different LLTMs [42]. In the present study, the R package eRm [49] was used for model specification and parameter estimation.

3.4. Model Specification

In this study, two LLTMs with the following variables were specified to estimate parameters α_k . In the first LLTM—called LLTM_{baseline}—it was coded to which of the seven skills each item belongs (i.e., dummy coding). This procedure mirrors the assumption that

there are specific cognitive demands to solve the items associated with each skill [23,50]. Hence, the assignment to the respective skills is assumed to sufficiently and totally explain the item difficulty in the LLTM_{baseline}.

The second LLTM—called LLTM_{extended}—additionally included parameters for the readability measures WSTF, LSTC, and GSTC, and the presence of visual representations described above. Hence, the LLTM_{extended} assumes that next to the scientific reasoning skills, the readability of text and the presence of visual representations also impose specific cognitive demands to process and encode information provided in the items, and to answer correctly [24,32,37,38].

4. Results

The Results Section is subdivided into three subsections: Basic Statistics, Descriptive Modeling, and Explanatory Modeling. The latter two sections refer to the two-step procedure of LLTM model evaluation, as described in Section 3.3.

4.1. Basic Statistics

Table 1 provides basic descriptive statistics and Pearson correlations for item difficulty and the variables considered in this study. Item difficulty was calculated as the proportion of correct responses (i.e., 1.0 = 100% correct responses). It is evident that the multiple-choice items had appropriate difficulty for the present sample, as about 47% of them were answered correctly ($M_{\text{ItemDiff}} = 0.47$). About 43% of the items contain a visual representation. Based on the WSTF and LSTC, the items would be considered rather easy to read. The LSTC is even higher than expected, indicating a very high local substantival textual cohesion. Only the average GSTC ($M_{\text{GSTC}} = 0.63$) indicates low global substantival textual cohesion of the items. Statistically significant correlations (i.e., $p < 0.05$) were only found between LSTC and GSTC ($r = 0.48$; medium effect size). Due to the medium effect size of this correlation, no serious problem of multicollinearity for further analysis occurs. Notably, no statistically significant correlations were found between item difficulty (ItemDiff) and the variables WSTF, LSTC, GSTC, and VisRep.

Table 1. Mean score (M), standard deviation (SD), and Pearson correlation coefficient (r) with related p -value for the respective variables. Expectance = values indicating appropriately understandable texts as suggested in [38]. ItemDiff = item difficulty; WSTF = 4. Wiener Sachtextformel; LSTC = local substantival textual cohesion; GSTC = global substantival textual cohesion; VisRep = item contains a visual representation (0 = no; 1 = yes).

	Expectance	$M \pm SD$		WSTF	LSTC	GSTC	VisRep
ItemDiff	—	0.47 ± 0.15	R	−0.28	−0.31	0.32	−0.20
			P	0.220	0.176	0.151	0.389
WSTF	5.4–8.4	6.43 ± 1.54	R		−0.15	−0.26	0.17
			P		0.502	0.251	0.468
LSTC	0.41–0.65	0.83 ± 0.38	R			0.48	0.34
			P			0.030	0.136
GSTC	0.70–0.89	0.63 ± 0.11	R				0.28
			P				0.221
VisRep	—	0.43 ± 0.51					

For further illustration, sample items can be found in Appendix A. These items represent the median score of WSTF ($M = 6.51$), LSTC ($M = 0.85$), and GSTC ($M = 0.61$), respectively.

Figure 1 below illustrates how the variables shown in Table 1 differ between the tasks for the seven skills of scientific reasoning. Kruskal–Wallis tests indicate significant differences between the skills for the variables GSTC ($H = 13.19$, $p = 0.040$) and VisRep ($H = 12.22$, $p = 0.045$). For GSTC, items related to the skills *planning investigations* ($M = 0.73$)

and *analyzing data and drawing conclusions* ($M = 0.78$) show rather high values, compared to lower values for the skills *formulating research questions* ($M = 0.53$), *generating hypotheses* ($M = 0.55$), *judging the purpose of models* ($M = 0.66$), *testing models* ($M = 0.54$), and *changing models* ($M = 0.64$). These five skills are below the suggested range of $0.70 < GSTC < 0.89$, unlike the others, indicating appropriately understandable texts in science education [39]. For VisRep, it is evident that items related to *formulating research questions*, *generating hypotheses*, and *planning investigations* do not contain visual representations, while most items related to the other skills do.

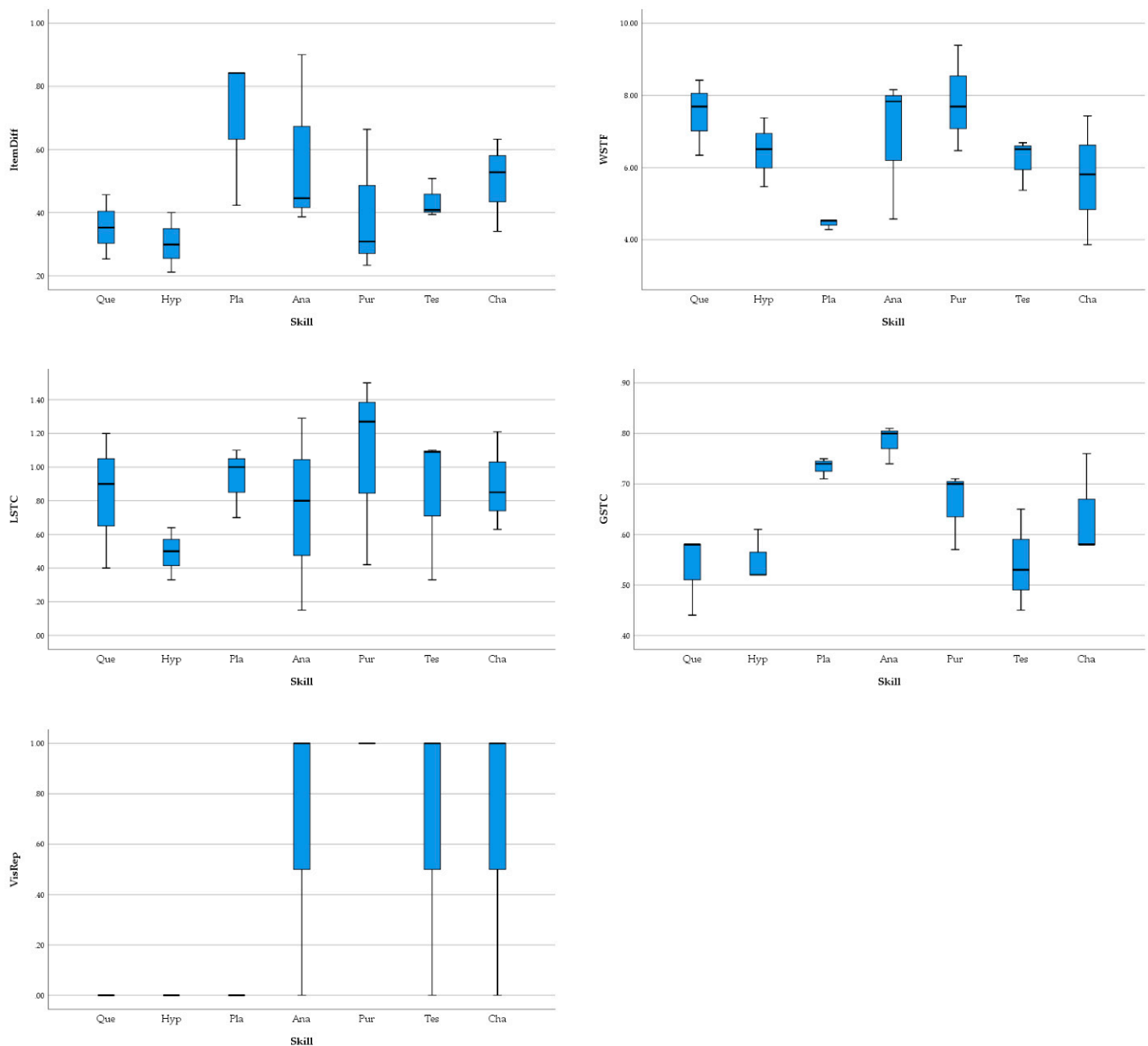


Figure 1. Boxplots for the variables ItemDiff (top left), WSTF (top right), LSTC (middle left), GSTC (middle right), and VisRep (bottom left) separated for the items assessing the seven skills *formulating research questions* (Que), *generating hypotheses* (Hyp), *planning investigations* (Pla), *analyzing data and drawing conclusions* (Ana), *judging the purpose of models* (Pur), *testing models* (Tes), and *changing models* (Cha).

4.2. Descriptive Rasch Modeling: One-Parameter Logistic Model (1PLM)

The fit between data and 1PLM has been evaluated and documented in previous studies in detail [2,5,16,21]. Here, MNSQs are reported, which indicates the discrepancy between the assumptions of the Rasch model and the data. MNSQ values are always positive because statistically, they are chi-square statistics divided by their degrees of freedom [51]. MNSQ values should lie in the range of 0.5–1.5 (“productive for measurement”) or 1.5–2.0 (“unproductive for construction of measurement but not degrading”), respectively, but not be >2.0 (“distorts or degrades the measurement system”) [48]. MNSQs can be calculated in two different versions—the outfit and the infit MNSQ. As the outfit MNSQ is more sensitive to outliers than the infit MNSQ, both statistics should be considered [51].

The MNSQ values in this study range between 0.7 and 1.2 (outfit MNSQ), and between 0.9 and 1.1 (infit MNSQ), respectively. Furthermore, the Andersen likelihood ratio test with the external split criterion “country” (i.e., Australia, Canada, UK) is not significant ($LR(40) = 46.22, p = 0.23$), thus indicating item homogeneity [49]. Person separation reliability is $rel. = 0.52$ and similar to previous reliability estimates for this instrument (e.g., [5]: EAP/PV reliability = 0.55; [16]: Cronbach’s Alpha = 0.60).

4.3. Explanatory Rasch Modeling: Linear Logistic Test Model (LLTM)

MNSQ values for both LLTMs indicate a reasonable fit between data and model (LLTM_{baseline}: $0.7 < \text{outfit MNSQ} < 1.6$; $0.7 < \text{infit MNSQ} < 1.5$; LLTM_{extended}: $0.5 < \text{outfit MNSQ} < 1.7$; $0.7 < \text{infit MNSQ} < 1.6$). Person separation reliability is $rel. = 0.46$ and 0.50 , respectively. Pearson correlations between the item parameters estimated in the LLTMs and the 1PLM are large for both the LLTM_{baseline} ($r = 0.65, p = 0.002$; i.e., $R^2 = 0.42$) and the LLTM_{extended} ($r = 0.86, p < 0.001$; i.e., $R^2 = 0.75$). The graphical model tests of the LLTMs and the 1PLM show that the item parameters scatter around the 45° line rather well for the LLTM_{extended}, while less so for the LLTM_{baseline} (Figure 2). This is also indicated by the empirical regression line (blue lines in Figure 2), which is closer to the 45° diagonal when comparing item difficulty parameters of the 1PLM and the LLTM_{extended} than when comparing these parameters of the 1PLM and the LLTM_{baseline}. In sum, the findings indicate that the item parameters estimated in the LLTM_{extended} were closer to the estimated parameters from the 1PLM, than those estimated in the LLTM_{baseline}.

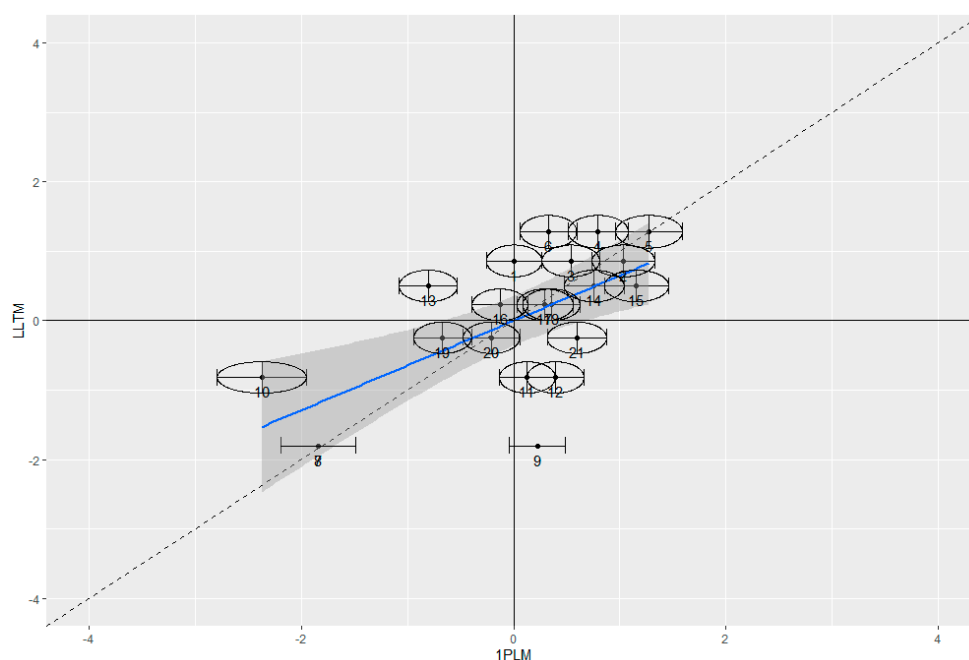


Figure 2. Cont.

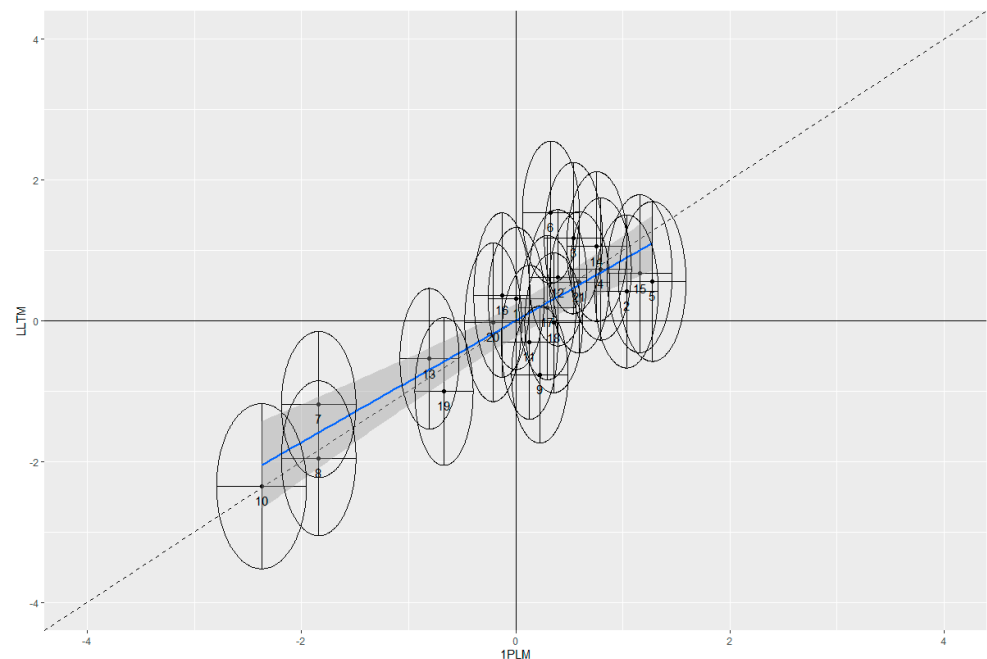


Figure 2. Graphical model tests comparing the 1PLM (x-axis) and the LLTM (y-axis) by the estimated item parameters (logits) for the LLTM_{baseline} (top) and the LLTM_{extended} (bottom). Each dot represents one item, with a 2*standard error of estimated item parameter (ellipses). The blue line is the empirical regression, with a 95% confidence interval in grey.

Table 2 provides the information criteria AIC and BIC and the log-likelihood difference test for model comparison between the 1PLM and the two LLTMs. AIC and BIC assess the relative model fit, with smaller values indicating the better fitting model. These values, therefore, indicate that the 1PLM fits better with the data than both LLTMs. The log-likelihood difference test also proposes a significantly better fit of the 1PL, compared to both LLTMs. Comparing both LLTMs, AIC and BIC indicate that the LLTM_{extended} fits better to the data than the LLTM_{baseline}.

Table 2. Model comparison between the 1PLM and both LLTMs (LogLik: marginal log-likelihood; AIC: Akaike information criterion; BIC: Bayesian information criterion; LD test: *p*-value of the log-likelihood difference test comparing the respective LLTM with the 1PLM).

Model	Parameter	LogLik	AIC	BIC	LD Test
1PLM	20	−3018	6076	6145	—
LLTM _{baseline}	6	−3282	6577	6597	<i>p</i> < 0.001
LLTM _{extended}	10	−3139	6299	6334	<i>p</i> < 0.001

Table 3 provides the α_k parameters as estimated in the two LLTMs. Positive α_k parameters indicate that the respective variable decreases item difficulty, while negative α_k parameters illustrate an increase in item difficulty. For the dummy coded variables representing the seven skills of scientific reasoning, *planning investigations* was chosen as the baseline because the related items ended up being rather easy (Figure 1). As the confidence intervals of most parameters in Table 3 do not include zero, they can be assumed to be significantly different from zero at the 5% level. Exceptions are WSTF, Pur, Test, and Cha in the LLTM_{extended}. Comparing the parameters in both LLTMs, it is evident the additional consideration of the variables WSTF, LSTC, GSTC, and VisRep reduces the effect of most of the dummy coded skills.

Table 3. Parameters estimated in the two LLTMs (SE = standard error; 95% CI = 95% confidence interval); lines with 95% CI including zero are formatted in grey.

Variable	α_k	SE	95% CI	
LLTM _{baseline}				
Que	−1.52	0.12	−1.75	−1.30
Hyp	−1.76	0.12	−2.00	−1.53
Ana	−0.56	0.11	−0.79	−0.34
Pur	−1.32	0.11	−1.54	−1.09
Tes	−1.16	0.11	−1.39	−0.94
Cha	−0.89	0.11	−1.11	−0.67
LLTM _{extended}				
WSTF	−0.04	0.03	−0.09	0.02
LSTC	−1.89	0.15	−2.17	−1.61
GSTC	5.61	0.78	4.09	7.14
VisRep	−0.79	0.11	−0.99	−0.58
Que	−0.53	0.19	−0.92	−0.15
Hyp	−1.57	0.16	−1.88	−1.26
Ana	−0.52	0.17	−0.85	−0.19
Pur	0.19	0.20	−0.20	0.58
Tes	0.28	0.20	−0.12	0.67
Cha	0.09	0.16	−0.22	0.41

In the LLTM_{extended}, the existence of visual representations ($\alpha_k = -0.79$) makes items harder to solve. Similarly, items related to the skills *formulating research questions*, *generating hypotheses*, and *analyzing data and drawing conclusions* are harder to solve than items related to the skill *planning investigations* (i.e., the baseline); this is also evident in Figure 1. As lower numbers of LSTC and GSTC are indicative of more complex texts, the α_k parameters of GSTC are in line with what was expected: the lower the GSTC is, the more difficult are the items to solve. Unlike expected, lower LSTC values decreased item difficulty ($\alpha_k = -1.89$).

As described above (Formula (5)), each item's difficulty is calculated in an LLTM as a linear (additive) combination of the item features' difficulty, with α_k as the estimated difficulty of item feature k . Based on the α_k values in Table 3, this means for the LLTM_{extended} that, for example, GSTC impacts item difficulty about seven times stronger than VisRep ($5.61/0.79 = 7.1$). It is important to note that α_k values are unstandardized and do not take the different scales of item features into account (e.g., binary variable VisRep vs. continuous variable GSTC).

5. Discussion

The purpose of this study was to investigate the effect of item features on item difficulty for a multiple-choice SRC assessment instrument established in science education [5,16,21]. More specifically, item features related to text complexity (4. Wiener Sachtextformel: WSTF; local and global substantival textual cohesion: LSTC and GSTC) and the presence of visual representations as figures or diagrams (i.e., VisRep) were investigated for their influence on item difficulty. The findings revealed that LSTC and GSTC, as well as VisRep, significantly impacted item difficulty in the multiple-choice assessment instrument, while WSTF did not. These findings are discussed below while acknowledging the limitations of this study.

In this study, the item features considered in the LLTM_{extended} explain about 75% of the variance in item difficulty estimated in the 1PLM—well above the limit of a large effect ($R^2 \geq 0.26$; [27]) and also higher than what has been found in similar studies (e.g., [28]:

$R^2 = 0.43$; [24]: $R^2 = 0.32$). Conversely, a variance explanation of 75% means that 25% of the variance in item difficulty estimated in 1PLM cannot be explained with the parameters specified in the LLTM_{extended} and might be attributable to individual differences. For example, general cognitive abilities such as verbal intelligence and problem-solving skills have been shown to significantly predict students' SRC [52].

The difference in variance explanation between the two LLTMs specified in this study suggests that 33% of the variance in item difficulty can be explained with the additional parameters related to text complexity and the existence of visual representations included in the LLTM_{extended}, that is, WSTF, LSTC, GSTC, and VisRep. The resulting amount of 33% is very similar to the result of an earlier study that found 32% [24] on item features affecting item difficulty in the German version of the instrument. This similarity in the effect of item features on item difficulty in both language versions of the instrument (English and German) is another indicator of test equivalence between the two versions [21].

A comparison of the parameters estimated in the LLTM_{baseline} and the LLTM_{extended} (Table 3) reveals that with the additional consideration of parameters related to text complexity and the presence of visual representations, the significant effect of *judging the purpose of models* (PUR), *testing models* (TES), and *changing models* (CHA), which were found in the LLTM_{baseline}, disappeared. This finding indicates that the significant effects of PUR, TES, and CHA, identified in the LLTM_{baseline}, might be artifacts caused by the effect of item features not considered in the LLTM_{baseline} and confounded with PUR, TES, and CHA. For example, all items related to PUR contain visual representations (Figure 1), while, on average, this applies to only 43% of the items (Table 1). Hence, the effect of PUR, identified in the LLTM_{baseline}, might have been caused by the presence of visual representations as figures or diagrams in the items related to PUR.

While the correlation analysis (Table 1) revealed no significant association between item difficulty and the item parameters of WSTF, LSTC, GSTC, and VisRep, these associations were found for most of the parameters in the LLTM_{extended}. This difference in findings is most likely caused by the fact that the correlation analysis was carried out based on the items (i.e., $N = 21$), a relatively small number to detect associations on a statistically significant level [26]. In contrast, the parameter estimation in the LLTM was performed based on a larger sample of individuals, or an $N = 243$ in this study.

Examining the individual parameters estimated in the LLTM_{extended} (Table 3), items containing visual representations tended to be harder to solve. This finding was also reported in [24] and described as unexpected, and potentially caused by the fact that visual representations in the items, "were often used to show complex scientific models and, hence, may increase the difficulty" (p. 8). Another explanation might be that the simultaneous encoding and processing of information provided in text and image can increase cognitive load and, hence, item difficulty [37]. As expected, lower global substantival textual cohesion increased item difficulty, with GSTC calculated as the proportion of substantives that appear more than once in a text (Formula (2)); however, unexpectedly, lower local substantival textual cohesion decreased item difficulty, with LSTC as the proportion of sentences with the same substantive as the preceding or subsequent sentence (Formula (3)). Both GSTC and LSTC measures are established indicators for text complexity and readability, with lower values indicating more difficult text [38]. The effect of GSTC on item difficulty most likely indicates that solving the items requires the encoding and processing of complex textual information provided in the item text globally, a task that is even more difficult with text that is challenging to read [24,32]. For the present multiple-choice items, this processing might involve respondents having to encode, process, and combine information that is textually presented in different parts of the item, such as the item stem and the answering options [50]. Hence, if information in the item stem and the answering options are more coherently presented (in terms of substantives), an item becomes easier to solve. For example, signal words, occurring both in the item stem and the attractor, can ease item difficulty [28]. The unexpected findings related to the effect of LSTC on item difficulty should be investigated further, for example, qualitatively, using cognitive interviews. One

plausible reason for the unexpected finding related to LSTC is that both GSTC and LSTC are typically used to analyze the readability of longer texts than what is included in the items of the present multiple-choice instrument [38]. Finally, the significant effects of some of the dummy coded skills (i.e., QUE, HYP, ANA; Table 3) illustrate that the items developed to assess the different skills of scientific reasoning require the application of specific procedural and epistemic knowledge to be solved [23].

The multiple-choice instrument under consideration in this study is already employed by scholars internationally in three language versions [2,16,21]. The findings of the present study shed light on specific cognitive demands that are necessary to correctly answering the items. These findings should be considered by scholars when interpreting test scores. Independent from the specific instrument, the study provides important insights about the influence of item features on item difficulty. These insights can inform the systematic development of a testing instrument that accounts for such features [27].

Naturally, this study has some limitations. The LLTM is well established for the analysis of item features and their influence on item difficulty within the approach of evaluating construct representation (e.g., [25,26]). Nevertheless, the assumption of an additive combination of the single features' difficulty, as described in Formula (5), is also criticized [43]. For example, a multiplicative combination of each item feature's influence on item difficulty might also be possible. Furthermore, in this study, only main effects were considered in LLTMs, but no interaction effects were considered between the specified variables. The variables considered in this study were also analyzed post hoc and were not systematically considered during item development; hence, the item features were not equally distributed across the items for the seven skills of SRC (e.g., items related to *formulating research questions*, *generating hypotheses*, and *planning investigations* do not contain visual representations at all; Figure 1). Finally, LLTMs assume that the specified item features completely (i.e., 100%) explain item difficulty [42], which was not the case in the present study. Despite a good explanation of item difficulty in the LLTM_{extended}, there is a significantly better model fit for the 1PLM (Table 2). The comparatively poor model fit of an LLTM is a common finding (e.g., [25,43]), which is explained with the strict assumption of a complete explanation of item difficulty by the specified item features [41]. The model comparison based on the information criteria, on the other hand, does not allow any statement about the absolute fit of the models considered [53]. Since a relatively worse model fit does not necessarily indicate an absolutely bad model fit, a check of the difficulty parameters estimated in the LLTM in the sense of a prognostic validation by replication studies is proposed [27,41]. This approach could be employed in the present context by developing additional items with systematically varied item features, followed by testing these features' influence on item difficulty again. Notwithstanding this issue of model fit, the comparison of the item difficulty parameters estimated in the 1PLM and both LLTMs allowed for an estimation of the amount of variance in item difficulty explained by the item features specified in the respective LLTM.

6. Conclusions

In this study, we investigated the effect of the item features WSTF, LSTC, GSTC, and VisRep on the difficulty of the items of a multiple-choice instrument to assess SRC in science education [5,21]. This analysis was based on the assumptions that the readability of text and the presence of visual representations impose specific cognitive demands to process and encode information provided in the items [24,32,37,38]. Furthermore, dummy-coded variables representing the specific skills of scientific reasoning were also considered in the analysis, assuming that specific cognitive demands (i.e., application of specific procedural and epistemic knowledge) are associated with each skill [23,50]. The findings illustrate that these variables, in sum, explain about 75% of the variance in item difficulty.

From a validity perspective, the similarity between the present findings and the previous study on the German version of the multiple-choice instrument [24] provides further evidence for test equivalence of both language versions [21]. From a cognitive point

of view [25], the findings of the present study suggest that specific cognitive demands are imposed by the readability of text and the presence of visual representations in multiple-choice assessment instruments. Specifically, the multiple-choice items analyzed in the present study appear to demand the encoding, processing, and combining of textually presented information from different parts of the items—such as item stem and answering options—while simultaneously encoding and processing information that is presented in both the text and visual representations. It has been shown that to solve the multiple-choice items used in this study, the application of procedural and epistemic knowledge is required [23,50]. The findings of this study illustrate that multiple-choice items on this assessment impose additional cognitive demands due to the necessity of processing text and visual representations.

Author Contributions: Conceptualization, M.K.; methodology, M.K.; investigation, M.K., S.K., and J.v.D.; resources, M.K., S.K., and J.v.D.; writing—original draft preparation, M.K.; writing—review and editing, M.K., S.K., and J.v.D.; visualization, M.K.; funding acquisition, M.K., S.K., and J.v.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the FUB Center for International Cooperation, Grant Number FMEEx2-2016-104, and the 2018 UBC-FUB Joint Funding Scheme, Grant Number FSP-2018-401.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Boards (or Ethics Committees) of The University of Melbourne (ID 11530, approved 3 January 2018), of the University of British Columbia (ID H18-01801, approved 23 July 2018), and of the University of Dundee (ID E2018-94, approved 15 July 2019).

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

Data Availability Statement: Data is available upon request to the first author.

Acknowledgments: The authors wish to thank Alexis Gonzalez and Song Xue in data collection and tabulation for the Canadian and UK samples, respectively, Christine Redman for her help with data collection for the Australian sample, and Jonna Kirchhof for her support in item analysis described in Section 3.2.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

The below items represent the median score of WSTF ($M = 6.51$), LSTC ($M = 0.85$), and GSTC ($M = 0.61$), respectively. Note that the items are presented in a tabular format for better reading and not in the same way as they appeared in the testing instrument. The attractor of each item is highlighted in italics.

Item “testing models 03” ($M_{WSTF} = 6.51$)

Item stem

Fraud with organic grocery bags?

Under the influence of oxygen, bacteria and fungi transform organic material mainly into carbon dioxide and water. This process of transformation is called composting. A part of the resulting substances is transformed into humus (dead organic soil matter). The following report was published in a newspaper: “The Deutsche Umwelthilfe (German Environmental Relief) launch accusations against two supermarket chains: The allegedly 100 % compostable grocery bags are not biodegradable at all; therefore they are just as ecologically harmful as common plastic bags.”

A team of experts has been asked to conduct a scientific investigation into how compostable are these organic grocery bags really?

Answering options

Which scientific question might underlie this investigation?

Tick one of the boxes below.

- What impact do the biological decomposition products from organic grocery bags have on the environment?
- What biological decomposition products are formed in the process of composting organic grocery bags?
- What materials comprise organic grocery bags?
- *Are there any substances formed in the process of composting organic grocery bags that cannot further be decomposed?*

Item “changing models 03” (MLSTC = 0.85)

Item stem

Language Acquisition

In physical reality, there is a variety of continuous transitions between different sounds, such as [ra] and [la]. While infants are aurally capable of perceiving all of these different transitions of sound, an imprint toward a specific language can be observed after the first year of life. Vocal expressions within different languages are then no longer perceived in their entirety but rather through a specific filter.

For this phenomenon of language acquisition, the following model was developed:

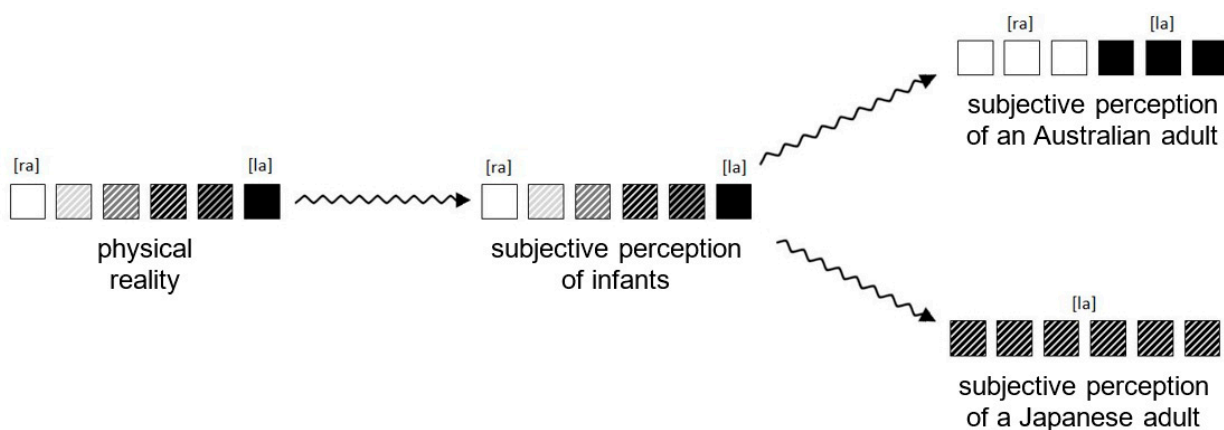


Figure. Model of language acquisition by sound perception.

The model predicts that Australians and Japanese acquire their language in different ways and the subjective perception of sounds develops differently.

Answering options

What reason would make it necessary to change the model?

Tick one of the boxes below.

The model has to be changed . . .

- . . . if the process of the subjective perception of [ra] and [la] in the language acquisition of English and Japanese is not explained.
- . . . if there are Japanese adults who learned English as a second language and have a distinct subjective perception of [ra] and [la].
- . . . if the subjective perception of [ra] and [la] cannot be applied to languages other than English and Japanese.
- . . . if there are Australian adults who do not have a distinct subjective perception of [ra] and [la].

Item “generating hypotheses 02” (M_{GSTC} = 0.61)

Item stem

In Outer Space

After many years of space missions, we know that existing conditions in space, such as zero gravity and cosmic radiation, harm the human body in the long run.

Previous stays in outer space were limited to a few months, whereas the scheduled flights to Mars will span many months—a considerably longer duration.

In a study, the health impacts of such long-lasting stays in outer space are to be investigated.

Answering options

Which scientific hypothesis might underlie this investigation?

Tick one of the boxes below.

- The human body needs additional protection against cosmic radiation during flights to outer space.
 - The human body shows little permanent damage from a short stay in outer space.
 - *The human body shows severe injuries when permanently being exposed to cosmic radiation.*
 - The existing conditions of zero gravity and radiation play a role in flights to Mars.
-

References

1. Erduran, S. Science education in the era of a pandemic: How can history, philosophy and sociology of science contribute to education for understanding and solving the Covid-19 crisis? *Sci. Educ.* **2020**, *29*, 233–235. [CrossRef]
2. Khan, S.; Krell, M. Scientific reasoning competencies: A case of preservice teacher education. *Can. J. Sci. Math. Technol. Educ.* **2019**, *19*, 446–464. [CrossRef]
3. Krell, M.; Dawborn-Gundlach, M.; van Driel, J. Scientific reasoning competencies in science teaching. *Teach. Sci.* **2020**, *66*, 32–42.
4. Mathesius, S.; Hartmann, S.; Upmeier zu Belzen, A.; Krüger, D. Scientific reasoning as an aspect of pre-service biology teacher education. In *The Future of Biology Education Research: Proceedings of the 10th Conference of European Researchers in Didactics of Biology (ERIDOB)*; Tal, T., Yarden, A., Eds.; Technion: Haifa, Israel, 2016; pp. 93–110.
5. Krell, M.; Redman, C.; Mathesius, S.; Krüger, D.; van Driel, J. Assessing pre-service science teachers' scientific reasoning competencies. *Res. Sci. Educ.* **2018**, *50*, 2305–2329. [CrossRef]
6. Lawson, A.E. The nature and development of scientific reasoning: A synthetic view. *Int. J. Sci. Math. Educ.* **2004**, *2*, 307–338. [CrossRef]
7. Morris, B.J.; Croker, S.; Masnick, A.M.; Zimmerman, C. The emergence of scientific reasoning. In *Current Topics in Children's Learning and Cognition*; Kloos, H., Morris, B.J., Amaral, J.L., Eds.; InTech: London, UK, 2012; pp. 61–82. [CrossRef]
8. Osborne, J. The 21st century challenge for science education: Assessing scientific reasoning. *Think. Ski. Creat.* **2013**, *10*, 265–279. [CrossRef]
9. European Commission. *Science Education for Responsible Citizenship*; European Commission: Luxembourg, 2015; Available online: <https://op.europa.eu/de/publication-detail/-/publication/a1d14fa0-8dbe-11e5-b8b7-01aa75ed71a1> (accessed on 25 August 2021).
10. OECD. *The High Cost of Low Educational Performance: The Long-Run Economic Impact of Improving PISA Outcomes (PISA)*; OECD: Paris, France, 2010; Available online: <https://www.oecd.org/pisa/44417824.pdf> (accessed on 25 August 2021).
11. Australian Curriculum, Assessment and Reporting Authority (ACARA). *The Australian Curriculum F-10: The Three Interrelated Strands of Science*; Australian Curriculum, Assessment and Reporting Authority: Sydney, Australia, 2018. Available online: <https://www.australiancurriculum.edu.au/f-10-curriculum/science/structure/> (accessed on 16 June 2020).
12. KMK. *Bildungsstandards im Fach Biologie Für Die Allgemeine Hochschulreife*; Wolters Kluwer: Hürth, Germany, 2020; Available online: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2020/2020_06_18-BildungsstandardsAHR_Biologie.pdf (accessed on 25 August 2021).
13. NGSS Lead States (Ed.) *Next Generation Science Standards: For States, by States*; The National Academies Press: Washington, DC, USA, 2013.
14. Osborne, J. Teaching Scientific Practices: Meeting the Challenge of Change. *J. Sci. Teach. Educ.* **2014**, *25*, 177–196. [CrossRef]
15. Göhner, M.; Krell, M. Preservice Science Teachers' Strategies in Scientific Reasoning: The Case of Modeling. *Res. Sci. Educ.* **2020**, *1*–20. [CrossRef]
16. Krüger, D.; Hartmann, S.; Nordmeier, V.; Upmeier zu Belzen, A. Measuring scientific reasoning competencies. In *Student Learning in German Higher Education*; Zlatkin-Troitschanskaia, O., Pant, H., Toepper, M., Lautenbach, C., Eds.; Springer: Wiesbaden, Germany, 2020; pp. 261–280. [CrossRef]
17. Stammen, A.; Malone, K.; Irving, K. Effects of modeling instruction professional development on biology teachers' scientific reasoning skills. *Educ. Sci.* **2018**, *8*, 119. [CrossRef]
18. Opitz, A.; Heene, M.; Fischer, F. Measuring scientific reasoning: A review of test instruments. *Educ. Res. Eval.* **2017**, *23*, 78–101. [CrossRef]
19. Ding, L.; Wei, X.; Mollohan, K. Does higher education improve student scientific reasoning skills? *Int. J. Sci. Math. Educ.* **2016**, *14*, 619–634. [CrossRef]
20. Hartmann, S.; Upmeier zu Belzen, A.; Krüger, D.; Pant, H.A. Scientific reasoning in higher education. *Z. Für Psychol.* **2015**, *223*, 47–53. [CrossRef]
21. Krell, M.; Mathesius, S.; van Driel, J.; Vergara, C.; Krüger, D. Assessing scientific reasoning competencies of pre-service science teachers: Translating a German multiple-choice instrument into English and Spanish. *Int. J. Sci. Educ.* **2020**, *42*, 2819–2841. [CrossRef]
22. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*; American Educational Research Association: Washington, DC, USA, 2014.

23. Mathesius, S.; Upmeier zu Belzen, A.; Krüger, D. Kompetenzen von Biologiestudierenden im Bereich der naturwissenschaftlichen Erkenntnisgewinnung: Entwicklung eines Testinstruments. *Erkenn. Biol.* **2014**, *13*, 73–88.
24. Stiller, J.; Hartmann, S.; Mathesius, S.; Straube, P.; Tiemann, R.; Nordmeier, V.; Krüger, D.; Upmeier zu Belzen, A. Assessing scientific reasoning: A comprehensive evaluation of item features that affect item difficulty. *Assess. Eval. High. Educ.* **2016**, *41*, 721–732. [CrossRef]
25. Baghaei, P.; Kubinger, K. Linear logistic test modeling with R. *Pract. Assess. Res. Eval.* **2015**, *20*, 1–11.
26. Embretson, S.; Daniel, R. Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychol. Sci. Q.* **2008**, *50*, 328–344.
27. Hartig, J.; Frey, A. Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychol. Rundsch.* **2012**, *63*, 43–49. [CrossRef]
28. Krell, M. Schwierigkeitserzeugende Aufgabenmerkmale bei Multiple-Choice-Aufgaben zur Experimentierkompetenz im Biologieunterricht: Eine Replikationsstudie. *Z. Für Didakt. Der Nat.* **2018**, *42*, 1–15. [CrossRef]
29. Embretson, S. Construct validity. *Psychol. Bull.* **1983**, *93*, 179–197.
30. Messick, S. Validity of psychological assessment. *Am. Psychol.* **1995**, *50*, 741–749. [CrossRef]
31. Schecker, H.; Neumann, K.; Theyßen, H.; Eickhorst, B.; Dickmann, M. Stufen experimenteller Kompetenz. *Z. Für Didakt. Der Nat.* **2016**, *22*, 197–213. [CrossRef]
32. Prenzel, M.; Häußler, P.; Rost, J.; Senkbeil, M. Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft* **2002**, *30*, 120–135.
33. Krey, O.; Schwanewedel, J. Lernen mit externen Repräsentationen. In *Theorien in Der Naturwissenschaftsdidaktischen Forschung*; Krüger, D., Parchmann, I., Schecker, H., Eds.; Springer: Cham, Germany, 2018; pp. 159–175.
34. Schnotz, W. Integrated Model of Text and Picture Comprehension. In *The Cambridge Handbook of Multimedia Learning*; Mayer, R., Ed.; Cambridge University Press: New York, NY, USA, 2005; pp. 72–103.
35. Wu, H.-K.; Puntambekar, S. Pedagogical affordances of multiple external representations in scientific processes. *J. Sci. Educ. Technol.* **2012**, *21*, 754–767. [CrossRef]
36. Schnotz, W.; Baadte, C. Surface and deep structures in graphics comprehension. *Mem. Cogn.* **2015**, *43*, 605–618. [CrossRef]
37. Paas, F.; Sweller, J. Implications of cognitive load theory for multimedia learning. In *The Cambridge Handbook of Multimedia Learning*; Mayer, R., Ed.; Cambridge University Press: New York, NY, USA, 2014; pp. 27–42.
38. Kulgemeyer, C.; Starauschek, E. Analyse der Verständlichkeit naturwissenschaftlicher Fachtexte. In *Methoden in Der Naturwissenschaftsdidaktischen Forschung*; Krüger, D., Parchmann, I., Schecker, H., Eds.; Springer: Berlin, Germany, 2014; pp. 241–253.
39. Bamberger, R.; Vanacek, E. *Lesen-Verstehen-Lernen-Schreiben: Die Schwierigkeitsstufen Von Texten in Deutscher Sprache*; Jugend und Volk: Wien, Austria, 1984.
40. Starauschek, E. Der Einfluss von Textkohäsion und gegenständlichen externen piktoralen Repräsentationen auf die Verständlichkeit von Texten zum Physiklernen. *Z. Für Didakt. Der Nat.* **2006**, *12*, 127–157.
41. Fischer, G. The linear logistic test model. In *Rasch Models*; Fischer, G., Molenaar, I., Eds.; Springer: New York, NY, USA, 1995; pp. 131–155.
42. Fischer, G.H. Linear Logistic Test Models. In *Encyclopedia of Social Measurement*; Kempf-Leonard, K., Ed.; Elsevier: Amsterdam, The Netherlands, 2005; pp. 505–514.
43. Hartig, J.; Frey, A.; Nold, G.; Klieme, E. An application of explanatory item response modeling for model-based proficiency scaling. *Educ. Psychol. Meas.* **2012**, *72*, 665–686. [CrossRef]
44. Embretson, S.; Reise, S. *Item Response Theory for Psychologists*; Erlbaum: Mahwah, NJ, USA, 2000.
45. Wilson, M.; de Boeck, P. Descriptive and explanatory item response models. In *Explanatory Item Response Models*; de Boeck, P., Wilson, M., Eds.; Springer: New York, NY, USA, 2004; pp. 43–74.
46. Wilson, M.; de Boeck, P.; Carstensen, C. Explanatory Item Response Models: A Brief Introduction. In *Assessment of Competencies in Educational Contexts*; Hartig, J., Klieme, E., Leutner, D., Eds.; Hogrefe Publishing: Göttingen, Germany, 2008; pp. 83–110.
47. Mair, P.; Hatzinger, R. Extended Rasch modeling. *J. Stat. Softw.* **2007**, *20*, 1–20. [CrossRef]
48. Wright, B.; Linacre, J. Reasonable mean-square fit values. *Rasch Meas. Trans.* **1994**, *8*, 370.
49. Mair, P.; Hatzinger, R.; Maier, M.J. *eRm: Extended Rasch Modeling*. 2020. Available online: <https://cran.r-project.org/package=eRm> (accessed on 25 August 2021).
50. Mathesius, S.; Upmeier zu Belzen, A.; Krüger, D. Eyetracking als Methode zur Untersuchung von Lösungsprozessen bei Multiple-Choice-Aufgaben zum wissenschaftlichen Denken. In *Lehr-und Lernforschung in Der Biologiedidaktik*; Hammann, M., Lindner, M., Eds.; Studienverlag: Innsbruck, Austria, 2018; pp. 225–244.
51. Linacre, J. What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Meas. Trans.* **2002**, *16*, 878.
52. Mathesius, S.; Krell, M.; Upmeier zu Belzen, A.; Krüger, D. Überprüfung eines Tests zum wissenschaftlichen Denken unter Berücksichtigung des Validitätskriteriums relations-to-other-variables. *Z. Für Pädagogik* **2019**, *65*, 492–510.
53. Burnham, K.; Anderson, D. Multimodel inference. *Sociol. Methods Res.* **2004**, *33*, 261–304. [CrossRef]

Article

Preservice Biology Teachers' Scientific Reasoning Skills and Beliefs about Nature of Science: How Do They Develop and Is There a Mutual Relationship during the Development?

Daniela Mahler ^{1,2,*} , Denise Bock ²  and Till Bruckermann ^{3,*} ¹ Biology Education, Freie Universität Berlin, 14195 Berlin, Germany² IPN—Leibniz Institute for Science and Mathematics Education, 24118 Kiel, Germany; bock@leibniz-ipn.de³ Institute of Education, Leibniz University Hannover, 30167 Hannover, Germany

* Correspondence: daniela.mahler@fu-berlin.de (D.M.); till.bruckermann@iew.uni-hannover.de (T.B.)

† Both authors contributed equally.

Abstract: Scientific reasoning (SR) skills and nature of science (NOS) beliefs represent important characteristics of biology teachers' professional competence. In particular, teacher education at university is formative for the professionalization of future teachers and is thus the focus of the current study. Our study aimed to examine the development of SR skills and NOS beliefs and their mutual relationship during teacher education. We applied paper-and-pencil tests to measure SR skills and NOS beliefs of 299 preservice biology teachers from 25 universities in Germany. The results of linear mixed models and planned comparisons revealed that both SR skills and NOS beliefs develop over the course of the study. Nevertheless, the development of SR skills and multiple aspects of NOS beliefs proceeds in different trajectories. Cross-lagged models showed a complex picture concerning the mutual relationship between SR skills and NOS beliefs during their development (both positive and negative). The current study contributes to the existing research because it is based on longitudinal data and allows—in contrast to cross-sectional research—conclusions about the *development* of SR skills and NOS beliefs.

Keywords: scientific reasoning; nature of science; preservice teachers; longitudinal study; cross-lagged panel

Citation: Mahler, D.; Bock, D.; Bruckermann, T. Preservice Biology Teachers' Scientific Reasoning Skills and Beliefs about Nature of Science: How Do They Develop and Is There a Mutual Relationship during the Development? *Educ. Sci.* **2021**, *11*, 558. <https://doi.org/10.3390/educsci11090558>

Academic Editors: Eila Jeronen, Moritz Krell, Andreas Vorholzer and Andreas Nehring

Received: 6 August 2021

Accepted: 13 September 2021

Published: 18 September 2021

Corrected: 30 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Fostering the scientific literacy of students is one of the core aims of science education in schools (e.g., [1] [Germany]; [2] [U.S.]). Science teachers' scientific reasoning (SR) skills and their beliefs about nature of science (NOS) represent key domains of science teachers' professional competence regarding science as inquiry. Science teachers with higher proficiency in SR skills are more likely to promote inquiry-based learning of their students [3,4]. Furthermore, science teachers need adequate NOS beliefs to integrate NOS teaching practices in their classrooms [5,6]. Accordingly, SR skills and NOS beliefs should be considered equally important as knowledge of other science concepts [7]. In different countries, standard documents of university teacher education, therefore, include SR skills and NOS beliefs (e.g., [8] [Germany]; [9] [U.S.]).

Because teacher education at the university is one of the most formative phases of the professionalization of teachers and the development of their professional competence [10], it should also be considered an important starting point for the development of SR skills and NOS beliefs. Previous research shows that preservice teachers' SR skills and NOS beliefs improve, at least to some degree, during teacher education at university, and that this development is related to appropriate learning opportunities provided in science education courses [11–13]. Regarding SR skills, preservice teachers are more skilled in graduate science education courses than students in graduate courses that did not explicitly

reflect upon scientific inquiry [12]. The more courses in university teacher education referred to NOS concepts, the higher is preservice teachers' understanding of NOS [11]. Nevertheless, empirical evidence based on longitudinal data allowing statements about the development of SR skills and NOS beliefs is rare (see [13], for an overview). In response to this desideratum, the first objective of the present study is to investigate the development of preservice science teachers' SR skills and NOS beliefs throughout university teacher education.

Although SR skills and NOS beliefs are two separate domains of teachers' professional competence, their interplay is important: Within teachers' professional competence, SR skills and NOS beliefs may be closely related because they both concern the development of scientific knowledge [14]. Whereas SR skills reflect preservice teachers' procedural knowledge about scientific inquiry processes (i.e., knowing how), NOS beliefs reflect their evaluations of how scientific knowledge comes into being (i.e., knowing why; [15]). Most research to date, however, was only able to correlate SR skills and NOS beliefs in cross-sectional designs [14,16]. Thus, the mutual relationship between SR skills and NOS beliefs and their development over time remains unexplored. For example, it is unclear whether NOS beliefs are beneficial for SR skills or vice versa (see [17], for an overview). Our second objective addresses this research gap by exploring the mutual relationship between SR skills and NOS beliefs during their development. With our results, we aim to contribute to understanding both constructs and their relationship, and the improvement in teacher education at university regarding SR skills and NOS beliefs.

1.1. Scientific Reasoning Skills and Beliefs about Nature of Science as Domains of Science Teachers' Professional Competence

To foster students' scientific literacy, science teachers need specific characteristics that are located within the concept of professional competence [18]. Professional competence covers different aspects of successful teachers' professional knowledge, motivational orientations, beliefs, values, and goals, and self-regulative skills [18]. These aspects are a critical resource for teachers to promote student learning [19–21]. One of the core assumptions of professional competence is that future teachers do not enter their careers with all of the desired characteristics but acquire them over time. This assumption implies that the aspects of teachers' professional competence are, in principle, learnable [22].

1.2. Scientific Reasoning Skills

SR skills are a procedural facet of teachers' content knowledge. Content knowledge represents the content-related domain of professional knowledge [18,23]. More specifically, SR skills represent the "knowing how" (i.e., about scientific inquiry processes; [15,24,25]). SR skills refer to an individual's ability to solve problems scientifically, that is, to a domain-specific set of knowledge and skills for scientific inquiry processes, which differ from domain-general cognitive strategies [26,27]. SR skills are different from domain-general cognitive abilities with which they only have medium but positive correlations [28]. SR skills comprise several subskills, such as formulating hypotheses, planning investigations, and analyzing and interpreting data. Furthermore, SR skills comprise different dimensions related to methods such as observing, investigations, and modeling [12,29,30]. The theoretical framework corresponding to the current study more precisely defines four subskills of SR (i.e., "formulating questions", "generating hypotheses", "planning investigations", "analyzing data and drawing conclusions") in the dimension of "conducting scientific investigations" [31] (p. 264). We refer to "skills" because they are—other than intelligence—trainable and also distinct from conceptual knowledge (see [32], for a similar account).

Regarding the assessment of SR skills, previous research noted a low validity of questionnaires based on ill-defined constructs of SR skills [25,33]. There have been few attempts to establish the psychometric quality of the questionnaires (e.g., [34]; see [32], for an overview). One notable exception was the development of a multiple-choice questionnaire for science teachers by theory-based item selection (see [31], for an overview), and testing of its validity [12,30] and applicability in different countries [29,35].

1.3. Beliefs about Nature of Science

Whereas scientific reasoning reflects the different activities in scientific inquiry (e.g., forming hypotheses, planning investigations, and analyzing and interpreting data), NOS reflects the epistemological basis of scientific inquiry and knowledge [5]. NOS beliefs, therefore, are different from knowledge of scientific inquiry. The location of NOS beliefs within the teachers' professional competence model reflects this conceptual difference [18], because NOS beliefs conceptually belong to teachers' beliefs [36]. Teachers' beliefs are defined as "psychologically held understandings and assumptions about phenomena or objects of the world that are felt to be true, have both implicit and explicit aspects and influence people's interactions with the world" [36] (p. 250). More precisely, NOS beliefs conceptually belong to teachers' epistemological beliefs related to the nature of knowledge or a particular science (see, [37] for mathematics education; see [38], for an overview). In science education, NOS beliefs reflect preservice teachers' evaluations of the characteristics of scientific knowledge and its production [5]. Despite the ongoing debate about the general aspects conceptualization of NOS, science education researchers, to a certain degree, agree on the inclusion of seven to ten aspects in the NOS conceptualization (i.e., the consensus view on aspects to be taught in schools; [39]). Previous research aligned the following aspects from different NOS conceptualizations: tentativeness; observations and inferences; creativity and imagination; subjectivity and objectivity; social and cultural embeddedness; diversity of scientific methods; and scientific theories and laws [39,40]. Recently, a comprehensive account of professional competence that teachers need for effective NOS instruction [38] added to the description of preservice teachers' NOS beliefs during teacher education [13] and provided a more prescriptive framework for teacher education.

The assessment of preservice teachers' NOS beliefs with questionnaires reflects the general aspects conceptualization of NOS. Questionnaires to assess NOS beliefs follow qualitative approaches, such as the Views of Nature of Science Questionnaire (VNOS; [41]), or quantitative Likert-type approaches, such as the questionnaire Student Understanding of Science and Scientific Inquiry (SUSSI; [42,43]). The Likert-type SUSSI questionnaire is based on the aspects from the VNOS [43]. Although Likert-type NOS questionnaires have been criticized [41], they are especially useful in research that assesses larger samples or repeatedly tests individuals and investigates the relationship between NOS beliefs and other constructs [44].

1.4. Development during Teacher Education

Kunter et al. [10] describe the first phase of teacher education at university as incredibly formative for the professionalization of teachers. Neumann and colleagues [45] provide a concise overview of the German teacher education system (in which our study is situated). Prospective teachers can choose from different teacher education programs to qualify for different school tracks (primary school, non-academic track, or academic track). Typically, they study two subjects. For prospective science teachers, it is essential to mention that they can study the separate science disciplines (biology, chemistry, and physics). On average, teacher education programs last five years (three years for the bachelor's phase and two years for the master's phase).

Both preservice teachers' SR skills and NOS beliefs profit from learning opportunities in science education courses during teacher education at university [11–13]. According to previous research, academic training generally appears to promote the development of SR skills [12,26,46]. Other research, however, suggests that the development of SR skills is more pronounced when courses promote explicit reflection on scientific inquiry [47–49]. Therefore, the development of preservice teachers' SR skills may vary throughout teacher education at university. In a cross-sectional study, university courses that require explicit reflection are part of the postgraduate phase of university teacher education [12]. The authors of this study assume that explicit reflection improved science teacher students' SR skills that were higher than those of natural science students, and they suggest further exploring this assumption in longitudinal studies. Adding to the cross-sectional findings,

one study provides evidence for the—at least moderate—development of SR skills during university teacher education in a longitudinal study on preservice teachers from two universities [31]. The development of SR skills was evident from four time points: the first and fourth semester of their undergraduate studies, and the first and fourth semester of their postgraduate studies (i.e., the 7th and 10th semesters in total).

Most research on preservice teachers' NOS beliefs stems from cross-sectional analysis and repeatedly shows that they do not possess what is considered to be adequate beliefs about NOS (e.g., [43]). Cross-sectional findings highlight that (future) science teachers often have an exclusively positive or idealistic image of science, even when the researchers accounted for the number of teaching years, the type of teacher education program, and the discipline (see [5], for an overview). Other research focused on how to promote adequate NOS beliefs and highlighted the effectiveness of explicit and reflective instruction [50–52]. Nonetheless, some aspects of NOS beliefs are more difficult to change than others (e.g., differences between scientific laws and theories; [52]). Less research focuses on how preservice teachers' NOS beliefs develop over time during university teacher education [11,53]. One study found a decline in adequate NOS beliefs in a sample of Turkish preservice teachers, although this study was cross-sectional [53]. Another study showed that adequate NOS beliefs increase with learning opportunities provided during university teacher education [11]. This study, however, also took a cross-sectional approach and did not consider how NOS beliefs of individual preservice teachers develop over time. Thus, longitudinal studies of preservice teachers' NOS beliefs are needed, particularly to explore how different aspects of NOS beliefs develop in relation to their difficulty [13].

1.5. The Interplay between SR Skills and NOS Beliefs

The two constructs may be related to each other because SR skills reflect the knowledge of how to pose questions scientifically, whereas NOS beliefs reflect the knowledge of why scientific inquiry proceeds in specific ways [14,54]. Both SR skills and NOS beliefs can be enhanced by appropriate instruction, such as explicit reflection about scientific inquiry [47,55]. Therefore, explicit teaching about scientific inquiry may lead to more appropriate NOS beliefs, for example, that theories are subject to change. Conversely, more appropriate NOS beliefs may promote more profound SR skills, such as drawing valid conclusions from data (see [16], for a similar account on the nature of scientific inquiry). Most research, however, either studied the development of SR skills (e.g., [12]) or NOS beliefs [11,53]. Other research that assessed both SR skills and NOS beliefs neither investigated their relationship (e.g., [56]) nor established a theoretical framework of how specific beliefs and skills may be related (e.g., [16]). Recently, the theoretical ScieNo-framework was developed [14]. In their framework, the authors assume that specific beliefs about the nature of scientific inquiry are related to specific SR skills: for example, SR skills for observations are related to views about the role of theory in observations [14,54]. Beyond the above-mentioned theoretical assumptions, there is also empirical evidence indicating a relationship between SR skills and NOS beliefs. The two studies that examined the relationship between SR skills and NOS beliefs [14,16], however, were cross-sectional, and they were not able to investigate the relation between two constructs during development. Furthermore, a longitudinal framework enables the investigation of the directions of the effects. Therefore, we suggest testing which NOS beliefs are related to SR skills in a longitudinal study.

1.6. The Current Research

In the current research, we investigated the development of SR skills and NOS beliefs in a longitudinal study with preservice biology teachers from German universities. Our study extends previous studies that investigated preservice teachers' SR skills (e.g., [12]) and NOS beliefs (e.g., [11,53]) only in cross-sectional designs and, therefore, were not able to describe the development using a longitudinal approach. In line with previous cross-sectional studies, we expected preservice teachers' SR skills and NOS beliefs to increase.

Furthermore, the current research investigated the mutual relationship between preservice teachers' SR skills and NOS beliefs during university teacher education. Our study extends previous research on the relationship between SR skills and NOS beliefs using a longitudinal approach to discern causal relationships between the constructs through a cross-lagged panel design. In line with previous studies that highlighted small to medium positive correlations between SR skills and NOS beliefs [14,16], we expected a positive mutual relationship between preservice teachers' SR skills and NOS beliefs. However, we did not assume a specific effect from one construct on another, because previous studies have only shown positive correlations. The following research questions guided our study:

1. How do preservice biology teachers' SR skills and their NOS beliefs develop over time during university teacher education?
2. How are preservice biology teachers' SR skills and NOS beliefs related to each other during the course of university teacher education?

2. Materials and Methods

2.1. Study Framework and Participants

This study was conducted in the longitudinal KeiLa project (Development of professional competence in science and mathematics teacher education). In KeiLa, preservice science and mathematics teachers from 25 universities in Germany attended up to four 4 h paper-and-pencil assessments between 2014 and 2017. The surveys took place independently of specific learning opportunities or courses. Instead, extra appointments were offered to participate in the study. We did this to obtain a general overview of the development of SR skills and NOS beliefs over the course of the study rather than to examine the effectiveness of specific learning opportunities.

In the current study, we refer to data of 299 preservice teachers (76% female; $M_{\text{age}} = 21.36$ years at first attendance, $SD_{\text{age}} = 2.59$). In the KeiLa project, a sequential-cohort design was conducted. We obtained annual data of preservice teachers enrolled in four consecutive semesters of semesters 1 to 11 throughout the four measurement points of the sequential-cohort design. All preservice teachers gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and no approval of the protocol by the local Ethics Committee was necessary. The reason for this is that the testing was carried out anonymously and proceeded in the familiar surroundings of university lecture halls, therefore causing no distress to the participating preservice teachers.

2.2. Instruments

2.2.1. Scientific Reasoning Skills

We assessed the SR skills of preservice biology teachers with 12 items developed in the Ko-WADiS project [12,31]. The single-choice items cover four subskills of SR with three items each: (1) formulating questions, (2) generating hypotheses, (3) planning investigations, and (4) analyzing data and drawing conclusions (see Table 1, for means and standard errors). We report a one-dimensional model based on dimensionality tests (see Section 2.3.2. Preliminary Analyses). The reliability of the scale is sufficient ($EAP/PV_{\text{Rel}} = 0.54$; based on concurrent calibration).

Table 1. Means and standard errors for scientific reasoning (SR) and nature of science (NOS) subscales of semesters 1 to 7.

Scale	1	3	5	7
	M^a (SE)	M (SE)	M (SE)	M (SE)
SR skills ^b	0.00 (0.07)	0.15 (0.08)	0.38 (0.07)	0.54 (0.08)
NOS beliefs ^c				
Observations and inferences	3.59 (0.05)	3.70 (0.06)	3.76 (0.05)	3.77 (0.06)
Tentativeness	3.84 (0.05)	3.89 (0.05)	3.94 (0.04)	4.10 (0.05)

Table 1. *Cont.*

Scale	1	3	5	7
	<i>M</i> ^a (SE)	<i>M</i> (SE)	<i>M</i> (SE)	<i>M</i> (SE)
Scientific theories and laws	2.67 (0.05)	2.74 (0.06)	2.80 (0.05)	2.84 (0.06)
Social/cultural embeddedness	3.46 (0.06)	3.65 (0.07)	3.72 (0.06)	3.80 (0.07)
Creativity and imagination	3.26 (0.06)	3.47 (0.07)	3.37 (0.06)	3.39 (0.07)
Scientific methods	3.67 (0.04)	3.71 (0.04)	3.74 (0.04)	3.87 (0.05)

^a Estimated marginal means and respective standard errors are based on linear mixed models; ^b Mean values of SR skills are based on WLE scores and can take any values centered around zero; ^c Mean values of NOS beliefs are based on Likert-type scales which range from 1 to 5.

2.2.2. Nature of Science Beliefs

We measured NOS beliefs with the “Student Understanding of Science and Scientific Inquiry” [42]. This contains 24 items that were assessed on 5-point Likert scales comprising 1 (*does not apply at all*), 2 (*does rather not apply*), 3 (*uncertain*), 4 (*largely applies*), and 5 (*fully applies*). The six NOS subscales (1) observations and inferences, (2) tentativeness, (3) scientific theories and laws, (4) social and cultural embeddedness, (5) creativity and imagination, and (6) scientific methods, were assessed with four items each (see Table 1, for means and standard errors). Ranges of the reliabilities (Cronbach’s α) of the subscales were as follows throughout the four semesters: from 0.45 to 0.68 for observations and inferences, from 0.50 to 0.59 for tentativeness, from 0.19 to 0.30 for scientific theories and laws, from 0.69 to 0.78 for social and cultural embeddedness, from 0.53 to 0.69 for creativity and imagination, and from 0.28 to 0.51 for scientific methods. They were calculated in R [57] with the “psych” package [58].

2.3. Analyses

2.3.1. Data Preparation

In our data set, we included participants from semesters 1 to 7 ($n_1 = 141$, $n_3 = 101$, $n_5 = 155$, $n_7 = 101$) because sample sizes in semesters 9 and 11 were too small for our analyses ($n_9 = 48$, $n_{11} = 8$). In our data set, preservice teachers were assigned to the respective semesters independent of the measurement points. Thereby, data were reshaped from the sequential-cohort design of the study to a longitudinal design.

2.3.2. Preliminary Analyses

In a first step, we conducted confirmatory factor analyses in Mplus [59] to check the assumed dimensionality of the constructs based on the subscales of the instruments (NOS beliefs: [42]; SR skills: [31]). Results revealed that a six-dimensional NOS model and a one-dimensional model of SR fitted the data significantly better than a one-dimensional model and a four-dimensional model, respectively (see Table 2).

Additionally, we calculated weighted likelihood estimation (WLE; [60]) scores for SR skills based on a one-parameter logistic item response theory model with concurrent calibration in R [57] using the “TAM” package [61].

Table 2. Chi-square difference ($\Delta\chi^2$), degrees of freedom difference (Δdf) and *p*-value of model comparison for one- and four-dimensional models of scientific reasoning (SR), and one- and six-dimensional models of nature of science (NOS) for semesters 1 to 7.

Semester	SR Skills			NOS Beliefs		
	$\Delta\chi^2$	Δdf	<i>p</i>	$\Delta\chi^2$	Δdf	<i>p</i>
1	7.44	6	0.283	181.44	15	<0.001
3	^a			161.61	15	<0.001
5	4.50	6	0.609	191.93	15	<0.001
7	10.11	6	0.120	142.83	15	<0.001

^a The four-dimensional model did not converge in semester 3.

2.3.3. Analyses concerning the Development

We chose a linear mixed model approach (LMM: [62]) to test if time (i.e., semesters) has a significant effect on preservice teachers' SR skills and NOS beliefs; that is, if SR skills and NOS beliefs develop over time. LMMs extend simple linear models by allowing both fixed and random effects. These models show various advantages. First, unlike in a repeated-measures analysis of variance, missing values can be easily handled (e.g., with restricted maximum likelihood [REML] estimation). Second, LMMs enable us to take the nested structure of our data (repeated observations nested in participants) into account. Thus, we can control for unobserved, time-invariant differences between participants. Third, LMMs allow us to control for specific autocorrelation structures, which can occur in repeated measures.

All LMMs were computed separately for each subscale of NOS beliefs and for SR skills. We fixed the correlations between time points to zero because previous checks revealed no critical autocorrelation structure to be considered. In our models, we treated the semester variable as a numeric fixed effect and the participants' ID as a grouping variable for the random effect, and we applied REML estimation. In addition to p -values, we computed the variance that is explained by all fixed effects (i.e., marginal R^2) and by fixed and random effects (i.e., conditional R^2) [63] because the trustworthiness of p -values provided for LMMs is the object of ongoing statistical discussions [64].

Finally, we examined planned comparisons of time points to further examine between which semesters significant mean changes for SR skills and NOS beliefs occur. First, we contrasted semesters 1 and 7 for SR skills and each subscale of NOS beliefs. In a second step, we compared consecutive semesters (i.e., semesters 1 vs. 3, 3 vs. 5, 5 vs. 7). p -values of the multiple comparisons were Bonferroni–Holm adjusted. We additionally calculated effect sizes (Cohen's d) based on the t -statistics for every comparison [65]. These are generally interpreted as small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$; [66]).

We used R [57] with the "nlme" package for LMMs [67], the "MuMIn" package for R^2 values [68], the "emmeans" package for planned comparisons [69], and the "effectsize" package for effect sizes d [70].

2.3.4. Analyses concerning the Mutual Relationship

We used a cross-lagged panel design with four waves (semesters 1, 3, 5, and 7) and specified the respective path models to examine the interactions of SR skills and NOS beliefs during teacher education. In the cross-lagged models, estimates of a later time point from one construct can directly be regressed on values of the previous time point from another construct (i.e., cross-lagged paths), and vice versa. Furthermore, we allowed parallel time points to be correlated and autoregressive paths. We computed a single model for each NOS subscale and its relationship with SR skills, including all four time points. Cross-lagged models were computed in R [57] with the "lavaan" package [71].

3. Results

3.1. Development of Scientific Reasoning Skills

Based on the linear mixed model (LMM), the semester has a significant effect on scientific reasoning values ($B = 0.09$, $SE = 0.02$, $t(198) = 5.43$, $p < 0.001$) with the marginal $R^2_m = 0.05$ and the conditional $R^2_c = 0.37$ (see Appendix A Table A1, for detailed LMM results).

The direct comparison of semester 1 with semester 7 shows that the mean of semester 7 is significantly higher than the mean of semester 1 (Estimate = 0.54, $SE = 0.11$, $t = 5.00$, $p < 0.001$, $d = 0.36$). Comparisons between sequential semesters yield no significant differences (Table 3; see Appendix A Table A2, for detailed comparison results).

Table 3. Estimates (Est.), standard errors (SE), and effect sizes (*d*) for subsequent time point comparisons of scientific reasoning and nature of science subscales.

Scale	1 vs. 7		1 vs. 3		3 vs. 5		5 vs. 7	
	Est. (SE)	<i>d</i>	Est. (SE)	<i>d</i>	Est. (SE)	<i>d</i>	Est. (SE)	<i>d</i>
Scientific reasoning skills	0.54 *** (0.11)	0.36	0.15 (0.10)	0.11	0.23 (0.10)	0.16	0.16 (0.10)	0.11
Nature of science beliefs								
Observations and inferenc.	0.17 * (0.07)	0.17	0.11 (0.06)	0.12	0.05 (0.06)	0.06	0.01 (0.06)	0.01
Tentativeness	0.26 *** (0.06)	0.28	0.05 (0.06)	0.06	0.05 (0.06)	0.06	0.16 * (0.06)	0.20
Scientific theories and laws	0.18 * (0.07)	0.17	0.07 (0.07)	0.08	0.06 (0.07)	0.06	0.04 (0.07)	0.05
Social and cultural embed.	0.34 *** (0.08)	0.29	0.19 * (0.08)	0.18	0.07 (0.08)	0.07	0.08 (0.07)	0.08
Creativity and imagination	0.14 (0.09)	0.11	0.22 * (0.08)	0.19	−0.10 (0.08)	−0.09	0.03 (0.08)	0.02
Scientific methods	0.20 ** (0.06)	0.23	0.05 (0.06)	0.06	0.03 (0.06)	0.04	0.12 (0.06)	0.15

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

3.2. Development of Nature of Science Beliefs

Based on the LMMs, the semester has a significant effect on the subscales observations and inferences ($B = 0.03$, $SE = 0.01$, $t(199) = 2.70$, $p = 0.008$, $R^2_m = 0.01$, $R^2_c = 0.59$), tentativeness ($B = 0.04$, $SE = 0.01$, $t(199) = 3.81$, $p < 0.001$, $R^2_m = 0.02$, $R^2_c = 0.58$), scientific laws and theories ($B = 0.03$, $SE = 0.01$, $t(199) = 2.60$, $p = 0.010$, $R^2_m = 0.01$, $R^2_c = 0.44$), social and cultural embeddedness ($B = 0.06$, $SE = 0.01$, $t(199) = 4.30$, $p < 0.001$, $R^2_m = 0.03$, $R^2_c = 0.53$), and scientific methods ($B = 0.03$, $SE = 0.01$, $t(199) = 3.07$, $p = 0.002$, $R^2_m = 0.01$, $R^2_c = 0.42$; see Appendix A Tables A3–A8, for detailed LMM results).

When comparing semester 1 and semester 7, we found a significant increase for five NOS subscales, that is, for observations and inferences (Estimate = 0.17, $SE = 0.07$, $t = 2.44$, $p = 0.011$, $d = 0.17$), tentativeness (Estimate = 0.26, $SE = 0.06$, $t = 4.00$, $p < 0.001$, $d = 0.28$), scientific laws and theories (Estimate = 0.17, $SE = 0.07$, $t = 2.41$, $p = 0.016$, $d = 0.17$), social and cultural embeddedness (Estimate = 0.34, $SE = 0.08$, $t = 4.12$, $p < 0.001$, $d = 0.29$), and scientific methods (Estimate = 0.20, $SE = 0.06$, $t = 3.22$, $p = 0.001$, $d = 0.23$). Sequential comparisons show that no consistent significant change from semester to semester occurs (Table 3; see Appendix A Tables A9–A14, for detailed comparison results).

3.3. The Mutual Relationship between SR Skills and NOS Beliefs

We found no significant cross-lagged paths between SR skills and the NOS subscales observations and inferences (all B s $< |0.13|$, all SE s < 0.16 , all p s > 0.05), scientific theories and laws (all B s $< |0.24|$, all SE s < 0.17 , all p s > 0.05), or creativity and imagination (all B s $< |0.15|$, all SE s < 0.15 , all p s > 0.05). However, we found significant relations in the other cross-lagged models, which are described in the following: A positive influence was found of NOS beliefs subscale tentativeness at semester 1 on SR skills at semester 3 ($B = 0.48$, $SE = 0.13$, $p < 0.001$). Another positive influence was found for social and cultural embeddedness beliefs at semester 5 on SR skills at semester 7 ($B = 0.30$, $SE = 0.12$, $p = 0.014$). Additionally, SR skills at semester 3 negatively influence scientific methods beliefs at semester 5 ($B = -0.22$, $SE = 0.07$, $p = 0.002$; see Figure 1).

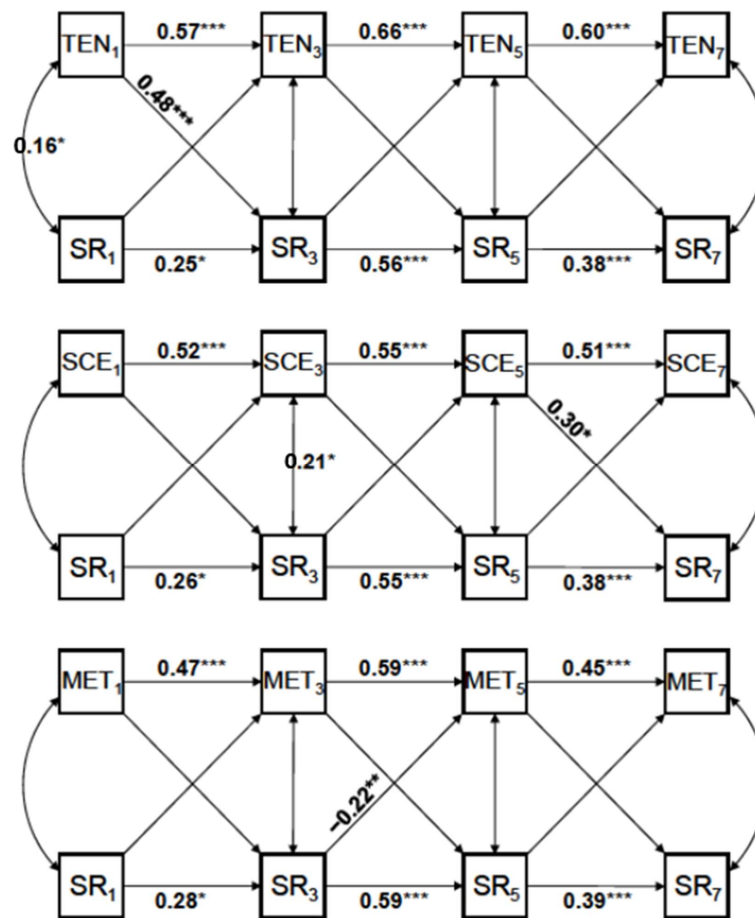


Figure 1. Unstandardized regression weights for (auto-)regression paths of cross-lagged models computed for scientific reasoning and nature of science subscales; tentativeness, TEN; social and cultural embeddedness, SCE; and scientific methods, MET; numbers in subscript refer to the respective semester; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4. Discussion

Science teachers' SR skills and NOS beliefs are essential characteristics when teaching science as inquiry [3–6]. To date, there are few longitudinal findings on either construct that allow making statements about their development during teacher education (e.g., SR skills: [31]). Thus, previous research left a gap concerning the development of individual preservice teachers over more extended periods of university teacher education [13]. Our study aimed to close this gap by taking a longitudinal approach to the development of preservice biology teachers' SR skills and NOS beliefs throughout university teacher education, and their mutual relationship during development. In line with evidence from cross-sectional studies, we assumed that both preservice teachers' SR skills (e.g., [12]), and NOS beliefs (e.g., [61]; cf. [62]), improve throughout teacher education at the university. Furthermore, we assumed a positive relationship between SR skills and NOS beliefs. However, we did not assume any direction in their relationship because previous findings were based on cross-sectional studies [14,16].

First, our results indicate that both preservice teachers' SR skills and NOS beliefs improved over the semesters of teacher education at university. The linear mixed models (LMM) revealed a positive impact of the semester variable on preservice teachers' SR skills, and on five of six NOS subscales, that is, observations and inferences, tentativeness, scientific theory and laws, social and cultural embeddedness, and scientific methods, but not creativity and imagination. Thus, our longitudinal study provides evidence that both develop throughout teacher education. When we account for the semesters in the

LMMs, they explain at least a slight variance (SR skills: 5% variance explained; NOS beliefs: 1–3% variance explained). We argue that the amount of explained variance appears plausible with regard to cross-sectional findings from other research. Our results support a previous study that found a comparable amount of variance of preservice teachers' NOS beliefs explained by semesters (i.e., 4%; [11]). Although our results indicate that preservice teachers' development of SR skills and NOS beliefs depends to some degree on their attendance of consecutive semesters, numerous other factors besides the semester remain unexplored. In this regard, the relatively high conditional R^2 values refer to a large amount of variance explained by time-invariant differences between participants, that is, differences that do not change over the considered period. These differences could be, for example, individual prerequisites such as a previously acquired degree (for SR skills: [35]) or the respective subject area of preservice teachers (for SR skills: [12]; for NOS beliefs: [51]).

Second, we explored when preservice teachers' SR skills and NOS beliefs develop throughout university teacher education by planned comparisons between the semesters 1 and 7 or 1 and 3, 3 and 5, and 5 and 7. For comparing semesters 1 and 7, we found that both SR skills and five of six aspects of NOS beliefs show a small to moderate increase in our study. Our results extend previous results from cross-sectional SR research [12,46] by using a longitudinal approach showing that preservice teachers' SR skills increase during semesters 1 to 7. The magnitude of this increase is comparable—at least at a descriptive level—to that in longitudinal data of semesters 1 to 7 from a previous study [31]. Although our results strengthen cross-sectional findings of the general development of NOS beliefs during teacher education (e.g., [11,52]; cf. [50,53]), most NOS beliefs' means at semester 7 still range between 3 (*uncertain*) and 4 (*largely applies*) on the Likert scale. The mean of the NOS subscale scientific theories and laws even remains below 3 throughout its development. Thus, we cannot assume that preservice teachers' development leads to *informed* views of NOS at the sample level [42,43,72].

Our more detailed analyses of the in-between semesters (1 vs. 3, 3 vs. 5, and 5 vs. 7), however, show that preservice teachers' SR skills and NOS beliefs developed in differing trajectories that are inconsistent in three ways: (1) they do not significantly improve during the in-between semesters, that is, for preservice teachers' SR skills and NOS beliefs about observations and inferences, scientific theories and laws, and scientific methods, only the comparison between semester 1 and 7 is significant; (2) some NOS beliefs do not steadily improve, such as tentativeness and social and cultural embeddedness, but show a significant increase with a small effect size ($d > 0.19$) only between two of the four consecutive semesters; (3) a slight decrease follows after initial positive development of NOS beliefs about creativity and imagination. These results suggest that the development does not just happen incidentally but that something must happen during teacher education that triggers this inconsistent picture of different trajectories. In principle, other studies also found that—at least for NOS beliefs—a decrease throughout teacher education is also possible [53]. Previous research highlighted that not all NOS aspects are equally changeable [13,52], which matches our result that only five aspects develop throughout teacher education. Our results complement prior research that showed that creativity and imagination, for example, are more likely to change [52], in that our results show that this belief changes mainly in the first few semesters and subsequently stagnates. Therefore, we assume that, in addition to the complexity of some NOS aspects [13] and individual differences among preservice teachers [50], learning opportunities in each semester should also be considered [11,12]. We assume that the uneven development of preservice teachers' SR skills and NOS beliefs depends on university teacher education's different learning opportunities. Previous research supports our assumption by indicating that learning opportunities are not equally distributed across teacher education at the university regarding pedagogical knowledge [73] and content knowledge, such as SR skills and NOS beliefs [11,12]. Thus, preservice teachers' SR skills and NOS beliefs are more likely to develop through multiple, interacting learning opportunities than through linearly cumulative learning opportunities (see [50], for a similar account). To further

understand the inconsistent picture, studies addressing learning opportunities during teacher education may help. A cross-sectional study found that the number of NOS-related learning opportunities is positively related to preservice biology teachers' NOS beliefs [11]. Both SR skills [12,74] and NOS beliefs [50–52] benefit from explicit and reflective learning opportunities. Accordingly, to further understand our results (e.g., why there are phases during the course of study that are more important for the development of SR skills and NOS beliefs compared to others), a closer examination of the number of learning opportunities, their distribution throughout teacher education, and their type (i.e., implicit vs. explicit) would have been helpful.

Third, we explored the mutual relationship between preservice teachers' SR skills and NOS beliefs during their development. We found that preservice teachers' NOS beliefs about tentativeness and social and cultural embeddedness positively influenced their SR skills: Less naïve NOS beliefs about tentativeness in semester 1 led to more profound SR skills in semester 3. Furthermore, less naïve NOS beliefs about the social and cultural embeddedness in semester 5 led to more profound SR skills in semester 7. Our longitudinal results align with previous research that established a positive correlation between NOS beliefs and SR skills in cross-sectional studies [14,16]. Furthermore, a mutual relationship between the SR skills and NOS beliefs appears plausible because both constructs refer to knowledge of scientific inquiry (knowing how and knowing why; [15]), and they can be improved through similar instructional approaches [47,55]. However, our longitudinal results also extend previous findings because our results revealed a much more inconsistent relationship that was limited to only some NOS beliefs and not stable across all semesters. The ScieNo-framework [14] may help to understand the inconsistent picture: skills for specific inquiry methods (such as conducting investigations and using models; i.e., dimensions of SR: [31]) are related to specific beliefs that are conceptually close (e.g., skills for observations and beliefs about the role of theory in observations). Therefore, we suggest that not all NOS beliefs are equally related to the SR skills of conducting investigations. In particular, NOS beliefs about tentativeness and the social and cultural embeddedness are challenging to grasp for preservice teachers [13,52], so that more adequate beliefs may have enhanced their SR skills in the following semesters. Other NOS beliefs that may be learned more easily probably do not positively influence SR skills. Future research should test our assumption that those NOS beliefs, in particular, that are more difficult to learn positively influence SR skills. Although we could separate six different NOS beliefs, we cannot relate them at this level of detail to different subskills of SR (e.g., formulating hypotheses) because we could not empirically separate the subscales for the SR skills. Furthermore, we used a short questionnaire that comprised 12 items of the dimension conducting scientific investigations from the whole item set that also includes the dimension of using models (see [31], for an overview). Thus, we suggest further research exploring the mutual relationship between different dimensions of SR skills, that is, for observing, experimenting, and modeling, and NOS beliefs in longitudinal studies.

Interestingly, we also found that preservice teachers' SR skills at semester 3 negatively influence their NOS beliefs about scientific methods at semester 5. Preservice teachers with more profound SR skills later had more naïve beliefs about scientific methods. We suspect that this negative effect may be related to a distortion in their beliefs about scientific methods when preservice teachers learn about methods of scientific inquiry in university teacher education. If a preservice teacher masters one inquiry method, such as how to conduct proper investigations, particularly well, this may lead to the idealistic (not appropriate) belief that this method is superior to the others. Our conjecture is in line with previous findings that show how naïve NOS beliefs develop with increasing study progress or Ph.D. degrees [53]. The authors explain this with the assumptions of Kuhn [75], who pointed out that during active engagement in research, the epistemological foundation fades into the background.

Furthermore, in university teacher education, science education courses have been shown to emphasize investigations, and particularly experiments, as teaching methods

that may promote preservice teachers' beliefs that there is only one scientific method [76]. Another explanation may be that limiting our SR questionnaire on the SR dimension of conducting investigations for test-economic reasons (see also [30]) may have led to a one-sided focus among the preservice biology teachers. Asking them only about conducting valid investigations probably made them believe that this was the only scientific method when they filled out the questionnaire on NOS beliefs. For future studies, we would recommend investigating such interactions between questionnaires on SR skills and NOS beliefs, and reflecting a greater variety of methods in the questionnaire on SR skills.

4.1. Strengths, Limitations, and Future Research

To the best of our knowledge, this study is the first to investigate the development of both preservice teachers' SR skills and NOS beliefs and their mutual relationship in a longitudinal design. More precisely, our study makes an essential contribution to the understanding of their development by using a longitudinal data set from 25 universities with adequate sample sizes and established instruments. Nevertheless, some limitations of this study should be discussed.

We used established and validated instruments for the assessment of both the SR skills and NOS beliefs (Ko-WADiS instrument: [31]; SUSSI instrument: [42]). Nevertheless, the reliabilities are partly in an unsatisfactory range in that they might have hindered us from detecting more substantial changes by the longitudinal design [50]. In comparison to previous research, the reliabilities determined in the current study are in the range of the typical values for the SUSSI, except for the subscale theories and laws (i.e., Cronbach's $\alpha = 0.44\text{--}0.89$: [42]; $\alpha = 0.16\text{--}0.86$: [77]) and for the Ko-WADiS instrument (i.e., EAP/PV reliability is: 0.54: [12]; 0.55: [34]).

The current study was designed to examine data from preservice teachers in semesters 1 to 11. Because the sample sizes were too small for semesters 9 and 11, these data had to be excluded from our analyses. Thus, we can only make statements for the bachelor's program that precedes the master's program in teacher education. We suggest for future research to examine the development of preservice teachers' SR skills and NOS beliefs during the master's program, because previous research suggests that the explicit learning opportunities that are particularly effective for both SR skills and NOS beliefs tend to occur in the latter part of teacher education [12].

The ScieNo-framework [14] helped us understand the mutual relationship between SR skills and NOS beliefs because it aligns the dimensions of SR skills, such as observing, experimenting, and modeling, with specific NOS beliefs. Unfortunately, we only used the 12 item short version of the Ko-WADiS instrument on the SR dimension when conducting investigations [78]. The full range of items includes another three SR skills of using models [31], so that the SR and NOS scales may be related to each other in a planned manner. In addition to test-economic reasons that led to the use of only one subscale for SR skills, it should be mentioned that the theoretical framework [14] was published after the current study was conducted between 2014 and 2017. However, our results highlight the mutual relationship between SR skills and the NOS beliefs about tentativeness, social and cultural embeddedness, and scientific methods. These mutual relationships are worth further investigation with more closely aligned instruments that are based on the ScieNo-framework.

4.2. Implications

Our results lead to implications, both for further research and for teacher education at university. They show that SR skills and multiple aspects of NOS beliefs develop differently; that is, some are easier to learn than others (e.g., [13]). The longitudinal study approach also suggests that preservice teachers' development of SR skills and NOS beliefs take different trajectories throughout teacher education, that is, the time point of development differs. As a next step, it would be essential to understand why SR skills and NOS beliefs take different trajectories during university teacher education. For this, the consideration of learning

opportunities is essential. Different studies suggest that explicit learning opportunities (i.e., learning opportunities that provide the opportunity for reflection) are particularly effective for developing SR skills and NOS beliefs; development of SR skills and NOS beliefs does not happen on the side. Accordingly, not only is the number of learning opportunities essential, but also their focus (implicit vs. explicit). In order to consider this, one could either refer to module manuals of teacher education or ask preservice teachers to report on the learning opportunities they had between two measurement points. The latter approach appears more promising because it provides information not only about the intended curriculum, but also about the implemented curriculum, that is, what actually took place [79]. We know from previous research that learning opportunities to explicitly reflect on scientific inquiry appear mainly in the master's program of German teacher education at the university [12]. Accordingly, it would be necessary to have an appropriate sample that covers both participants in the bachelor's program and the master's program.

Our results on the relationship between SR skills and NOS beliefs show an inconsistent picture in that there is no mutual relationship between all aspects of NOS beliefs and SR skills. To learn more about their mutual relationship, the ScieNo-framework [14] can be consulted to derive and further investigate hypotheses regarding the relationship between specific aspects of NOS beliefs and SR skills; for example, SR skills for observations are related to views about the role of theory in observations. A longitudinal approach—as in this study—would offer two advantages. First, it subjects the framework to empirical testing in a longitudinal perspective. Second, it would allow us to examine the extent to which learning opportunities for SR skills are also conducive to NOS beliefs, and vice versa. However, careful planning is necessary for such an investigation. For example, when selecting the instruments, it must be ensured that the constructs can be combined at an appropriate level of detail.

Our results also provide suggestions for improving teacher education. Our results on the different trajectories of preservice teachers' SR skills and NOS beliefs are significant in this regard. Our results make clear that preservice teachers' SR skills and NOS beliefs do not simply co-evolve but that their mutual relationship is much more complex across teacher education at university. The positive influence of specific aspects of preservice teachers' NOS beliefs on their SR skills depends on how difficult certain NOS aspects are to learn and when they develop due to learning opportunities in teacher education. Accordingly, a blanket consideration in teacher education is not expedient because inquiry-based learning does not automatically lead to the development of SR skills and NOS beliefs. Instead, learning opportunities must be created that explicitly relate the corresponding SR skills and NOS beliefs to each other (ScieNo-framework; [14]) and ideally provide space for reflection on this interplay.

Furthermore, we found that the more profound SR skills preservice teachers had in semester 3, the less informed were their NOS beliefs about scientific methods in semester 5. We suggest that this might stem from the negative impact of a bias concerning a single scientific method. It is possible that a strong focus on conducting scientific inquiry in undergraduate studies (i.e., doing science: [80]; e.g., [12]), and in our test, leads preservice teachers to idealistic but inadequate beliefs about methods in scientific research. Thus, teacher education should reflect the broad repertoire of methods in scientific research and provide opportunities for reflection on the use of those methods.

5. Conclusions

We investigated the development of SR skills and NOS beliefs—two characteristics of effective science teachers—and their mutual relationship during the undergraduate studies of teacher education at university. Our results add to previous research by taking a longitudinal approach to show how SR skills and NOS beliefs develop throughout teacher education. We present evidence for differing trajectories in the development of SR skills and multiple aspects of NOS beliefs that hints at the importance of further investigation of the learning opportunities. The present findings do not yet account for

learning opportunities in university teacher education, so it would be interesting to see whether the trajectories vary in university teacher education of other countries. If patterns emerge in the country comparison, the different learning opportunities leading to these trajectories can be examined in more detail. Furthermore, we present evidence that the mutual relationship between SR skills and NOS beliefs is stronger for specific aspects of preservice teachers' NOS beliefs and specific time points in their development. Thus, during teacher education at university, preservice teachers' SR skills and NOS beliefs are intertwined in their development, but further research is needed to truly understand their interplay and dependence on learning opportunities.

Author Contributions: Conceptualization, D.M. and T.B.; formal analysis, D.B.; data curation, D.M. and D.B.; writing—original draft preparation, D.M., D.B. and T.B.; writing—review and editing, D.M. and T.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Leibniz Association under grant number SAW-2014-IPN-1.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and no approval of the protocol by the local Ethics Committee was necessary. The reason for this is that the testing was carried out anonymously and proceeded in the familiar surroundings of university lecture halls, therefore causing no distress to the participating preservice teachers.

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 [The study is ongoing].

Acknowledgments: We thank Ute Harms for project administration and funding acquisition for the study at hand.

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

Appendix A

Table A1. Outcomes of the linear mixed model and effect sizes (marginal R^2_m for fixed effects, conditional R^2_c for fixed and random effects) for *scientific reasoning* with semester as a fixed effect and participants as a random effect.

Random Effects	Variance	SD			
Participant (Intercept)	0.26	0.51			
Residual	0.50	0.71			
Fixed Effects	B	SE	df	t	p
Intercept	0.00	0.07	293	−0.07	0.943
Semester	0.09	0.02	198	5.43	<0.001
Effect Size	R^2_m	R^2_c			
Semester	0.05	0.37			

Table A2. Estimates, standard errors (SE), t-values, p-values, and effect sizes (d) for planned comparisons of *scientific reasoning*.

Parameter	1 vs. 7	1 vs. 3	3 vs. 5	5 vs. 7
Estimate	0.54	0.15	0.23	0.16
SE	0.11	0.10	0.10	0.10
t	5.00	1.50	2.30	1.57
p	<0.001	0.235	0.067	0.235
d [95% CI]	0.36 [0.21, 0.50]	0.11 [−0.03, 0.25]	0.16 [0.02, 0.31]	0.11 [−0.03, 0.35]

Table A3. Outcomes of the linear mixed model and effect sizes (marginal R^2_m for fixed effects, conditional R^2_c for fixed and random effects) for *observations and inferences* with semester as a fixed effect and participants as a random effect.

Random Effects	Variance	SD			
Participant (Intercept)	0.25	0.50			
Residual	0.18	0.43			
Fixed Effects	B	SE	df	t	p
Intercept	3.61	0.05	295	75.68	<0.001
Semester	0.03	0.01	199	2.70	0.008
Effect Size	R^2_m	R^2_c			
Semester	0.01	0.59			

Table A4. Outcomes of the linear mixed model and effect sizes (marginal R^2_m for fixed effects, conditional R^2_c for fixed and random effects) for *tentativeness* with semester as a fixed effect and participants as a random effect.

Random Effects	Variance	SD			
Participant (Intercept)	0.20	0.45			
Residual	0.15	0.39			
Fixed Effects	B	SE	df	t	p
Intercept	3.82	0.04	294	88.59	<0.001
Semester	0.04	0.01	199	3.81	<0.001
Effect Size	R^2_m	R^2_c			
Semester	0.02	0.58			

Table A5. Outcomes of the linear mixed model and effect sizes (marginal R^2_m for fixed effects, conditional R^2_c for fixed and random effects) for *scientific theories and laws* with semester as a fixed effect and participants as a random effect.

Random Effects	Variance	SD			
Participant (Intercept)	0.17	0.41			
Residual	0.22	0.46			
Fixed Effects	B	SE	df	t	p
Intercept	2.67	0.05	294	58.24	<0.001
Semester	0.03	0.01	199	2.60	0.010
Effect Size	R^2_m	R^2_c			
Semester	0.01	0.44			

Table A6. Outcomes of the linear mixed model and effect sizes (marginal R^2_m for fixed effects, conditional R^2_c for fixed and random effects) for *social and cultural embeddedness* with semester as a fixed effect and participants as a random effect.

Random Effects	Variance	SD			
Participant (Intercept)	0.29	0.53			
Residual	0.27	0.51			
Fixed Effects	B	SE	df	t	p
Intercept	3.48	0.05	294	63.80	<0.001
Semester	0.06	0.01	199	4.30	<0.001
Effect Size	R^2_m	R^2_c			
Semester	0.03	0.53			

Table A7. Outcomes of the linear mixed model and effect sizes (marginal R^2_m for fixed effects, conditional R^2_c for fixed and random effects) for *creativity and imagination* with semester as a fixed effect and participants as a random effect.

Random Effects	Variance	SD			
Participant (Intercept)	0.21	0.46			
Residual	0.33	0.57			
Fixed Effects	B	SE	df	t	p
Intercept	3.30	0.05	294	60.34	<0.001
Semester	0.02	0.01	199	1.40	0.164
Effect Size	R^2_m	R^2_c			
Semester	0.00	0.40			

Table A8. Outcomes of the linear mixed model and effect sizes (marginal R^2_m for fixed effects, conditional R^2_c for fixed and random effects) for *scientific methods* with semester as a fixed effect and participants as a random effect.

Random Effects	Variance	SD			
Participant (Intercept)	0.11	0.33			
Residual	0.16	0.40			
Fixed Effects	B	SE	df	t	p
Intercept	3.66	0.04	294	95.05	<0.001
Semester	0.03	0.01	199	3.07	0.002
Effect Size	R^2_m	R^2_c			
Semester	0.02	0.42			

Table A9. Estimates, standard errors, *t*-values, *p*-values, and effect sizes (*d*) for planned comparisons of *observations and inferences*.

Parameter	1 vs. 7	1 vs. 3	3 vs. 5	5 vs. 7
Estimate	0.17	0.11	0.05	0.01
SE	0.07	0.06	0.06	0.06
<i>t</i>	2.44	1.71	0.85	0.15
<i>p</i>	0.015	0.269	0.796	0.877
<i>d</i> [95% CI]	0.17 [0.03, 0.31]	0.12 [−0.02, 0.26]	0.06 [−0.08, 0.20]	0.01 [−0.13, 0.15]

Table A10. Estimates, standard errors, *t*-values, *p*-values, and effect sizes (*d*) for planned comparisons of *tentativeness*.

Parameter	1 vs. 7	1 vs. 3	3 vs. 5	5 vs. 7
Estimate	0.26	0.05	0.05	0.16
SE	0.06	0.06	0.06	0.06
<i>t</i>	4.00	0.81	0.86	2.76
<i>p</i>	<0.001	0.779	0.779	0.019
<i>d</i> [95% CI]	0.28 [0.14, 0.43]	0.06 [−0.08, 0.20]	0.06 [−0.08, 0.20]	0.20 [0.06, 0.34]

Table A11. Estimates, standard errors, *t*-values, *p*-values, and effect sizes (*d*) for planned comparisons of scientific theories and laws.

Parameter	1 vs. 7	1 vs. 3	3 vs. 5	5 vs. 7
Estimate	0.18	0.07	0.06	0.04
SE	0.07	0.07	0.07	0.07
<i>t</i>	2.41	1.06	0.90	0.65
<i>p</i>	0.017	0.873	0.873	0.873
<i>d</i> [95% CI]	0.17 [0.03, 0.31]	0.08 [−0.06, 0.22]	0.06 [−0.08, 0.20]	0.05 [−0.09, 0.19]

Table A12. Estimates, standard errors, *t*-values, *p*-values, and effect sizes (*d*) for planned comparisons of social and cultural embeddedness.

Parameter	1 vs. 7	1 vs. 3	3 vs. 5	5 vs. 7
Estimate	0.34	0.19	0.07	0.08
SE	0.08	0.08	0.08	0.07
<i>t</i>	4.11	2.53	0.92	1.07
<i>p</i>	<0.001	0.037	0.571	0.571
<i>d</i> [95% CI]	0.29 [0.15, 0.43]	0.18 [0.04, 0.32]	0.07 [−0.07, 0.21]	0.08 [−0.06, 0.22]

Table A13. Estimates, standard errors, *t*-values, *p*-values, and effect sizes (*d*) for planned comparisons of creativity and imagination.

Parameter	1 vs. 7	1 vs. 3	3 vs. 5	5 vs. 7
Estimate	0.14	0.22	−0.10	0.03
SE	0.09	0.08	0.08	0.08
<i>t</i>	1.60	2.68	−1.27	0.31
<i>p</i>	0.112	0.024	0.412	0.755
<i>d</i> [95% CI]	0.11 [−0.03, 0.25]	0.19 [0.05, 0.33]	−0.09 [−0.23, 0.05]	0.02 [−0.12, 0.16]

Table A14. Estimates, standard errors, *t*-values, *p*-values, and effect sizes (*d*) for planned comparisons of scientific methods.

Parameter	1 vs. 7	1 vs. 3	3 vs. 5	5 vs. 7
Estimate	0.20	0.05	0.03	0.12
SE	0.06	0.06	0.06	0.06
<i>t</i>	3.22	0.81	0.52	2.16
<i>p</i>	0.002	0.842	0.842	0.096
<i>d</i> [95% CI]	0.23 [0.09, 0.37]	0.06 [−0.08, 0.20]	0.04 [−0.10, 0.18]	0.15 [0.01, 0.29]

References

1. Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss [National Educational Standards for the Intermediate School Leaving Certificate in Biology]*; Luchterhand: München, Germany, 2005.
2. NGSS Lead States. *Next Generation Science Standards: For States, by States*; The National Academies Press: Washington, DC, USA, 2013.
3. Capps, D.K.; Crawford, B.A.; Constat, M.A. A review of empirical literature on inquiry professional development: Alignment with best practices and a critique of the findings. *J. Sci. Teacher Educ.* **2012**, *23*, 291–318. [CrossRef]
4. Schwarz, C. Developing preservice elementary teachers' knowledge and practices through modeling-centered scientific inquiry. *Sci. Educ.* **2009**, *93*, 720–744. [CrossRef]
5. Lederman, N.G.; Lederman, J.S. Research on Teaching and Learning of Nature of Science. In *Handbook of Research on Science Education*; Lederman, N.G., Abell, S.K., Eds.; Routledge: New York, NY, USA, 2014; ISBN 9780203097267.
6. Capps, D.K.; Crawford, B.A. Inquiry-based professional development: What does it take to support teachers in learning about inquiry and nature of science? *Int. J. Sci. Educ.* **2013**, *35*, 1947–1978. [CrossRef]



7. Glaze, A. Teaching and Learning Science in the 21st Century: Challenging Critical Assumptions in Post-Secondary Science. *Educ. Sci.* **2018**, *8*, 12. [CrossRef]
8. Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung (Beschluss der Kultusministerkonferenz vom 16.10.2008 i. d. F. vom 16.05.2019). 2019. Available online: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf (accessed on 7 September 2021).
9. National Science Teacher Association. NSTA Standards for Science Teacher Preparation. Available online: <https://www.nsta.org/nsta-standards-science-teacher-preparation> (accessed on 7 September 2021).
10. Kunter, M.; Kleickmann, T.; Klusmann, U.; Richter, D. The Development of Teachers' Professional Competence. In *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers: Results from the COACTIV Project*; Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., Neubrand, M., Eds.; Springer: New York, NY, USA, 2013; ISBN 978-1-4614-5148-8.
11. Bruckermann, T.; Ochsen, F.; Mahler, D. Learning opportunities in biology teacher education contribute to understanding of nature of science. *Educ. Sci.* **2018**, *8*, 103. [CrossRef]
12. Hartmann, S.; Upmeyer zu Belzen, A.; Krüger, D.; Pant, H.A. Scientific Reasoning in Higher Education. *Z. Psychol.* **2015**, *223*, 47–53. [CrossRef]
13. Cofré, H.; Núñez, P.; Santibáñez, D.; Pavez, J.M.; Valencia, M.; Vergara, C. A Critical Review of Students' and Teachers' Understandings of Nature of Science. *Sci. Educ.* **2019**, *28*, 205–248. [CrossRef]
14. Reith, M.; Nehring, A. Scientific reasoning and views on the nature of scientific inquiry: Testing a new framework to understand and model epistemic cognition in science. *Int. J. Sci. Educ.* **2020**, *42*, 2716–2741. [CrossRef]
15. Kind, P.; Osborne, J. Styles of scientific reasoning: A cultural rationale for science education? *Sci. Educ.* **2017**, *101*, 8–31. [CrossRef]
16. Kremer, K.; Specht, C.; Urhahne, D.; Mayer, J. The relationship in biology between the nature of science and scientific inquiry. *J. Biol. Educ.* **2013**, *48*, 1–8. [CrossRef]
17. Rönnebeck, S.; Bernholt, S.; Ropohl, M. Searching for a common ground—A literature review of empirical research on scientific inquiry activities. *Stud. Sci. Educ.* **2016**, *52*, 161–197. [CrossRef]
18. Baumert, J.; Kunter, M. The COACTIV Model of Teachers' Professional Competence. In *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers: Results from the COACTIV Project*; Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., Neubrand, M., Eds.; Springer: New York, NY, USA, 2013; ISBN 978-1-4614-5148-8.
19. Baumert, J.; Kunter, M.; Blum, W.; Brunner, M.; Voss, T.; Jordan, A.; Klusmann, U.; Krauss, S.; Neubrand, M.; Tsai, Y.-M. Teachers' Mathematical Knowledge, Cognitive Activation in the Classroom, and Student Progress. *Am. Educ. Res. J.* **2010**, *47*, 133–180. [CrossRef]
20. Brigham, F.J.; Scruggs, T.E.; Mastropieri, M.A. Teacher Enthusiasm in Learning Disabilities Classrooms: Effects on Learning and Behavior. *Learn. Disabil. Res. Pract.* **1992**, *7*, 68–73.
21. Mahler, D.; Großschedl, J.; Harms, U. Using doubly latent multilevel analysis to elucidate relationships between science teachers' professional knowledge and students' performance. *Int. J. Sci. Educ.* **2017**, *39*, 213–237. [CrossRef]
22. Weinert, F.E. Concept of competence: A conceptual clarification. In *Defining and Selecting Key Competencies*; Rychen, D.S., Salganik, L.H., Eds.; Hogrefe & Huber: Seattle, WA, USA, 2001; pp. 45–65.
23. Shulman, L.S. Those Who Understand: Knowledge Growth in Teaching. *Educ. Res.* **1986**, *15*, 4–14. [CrossRef]
24. Gut-Glanzmann, C.; Mayer, J. Experimentelle Kompetenz. In *Theorien in der Naturwissenschaftsdidaktischen Forschung*; Krüger, D., Parchmann, I., Schecker, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; pp. 121–140, ISBN 978-3-662-56319-9.
25. Osborne, J. The 21st century challenge for science education: Assessing scientific reasoning. *Think. Ski. Creat.* **2013**, *10*, 265–279. [CrossRef]
26. Lawson, A.E.; Clark, B.; Cramer-Meldrum, E.; Falconer, K.A.; Sequist, J.M.; Kwon, Y.-J. Development of Scientific Reasoning in College Biology: Do Two Levels of General Hypothesis-Testing Skills Exist? *J. Res. Sci. Teach.* **2000**, *37*, 81–101. [CrossRef]
27. Bao, L.; Cai, T.; Koenig, K.; Fang, K.; Han, J.; Wang, J.; Liu, Q.; Ding, L.; Cui, L.; Luo, Y.; et al. Learning and scientific reasoning. *Science* **2009**, *323*, 586–587. [CrossRef] [PubMed]
28. Mayer, D.; Sodian, B.; Koerber, S.; Schwippert, K. Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learn. Instr.* **2014**, *29*, 43–55. [CrossRef]
29. Krell, M.; Mathesius, S.; van Driel, J.; Vergara, C.; Krüger, D. Assessing scientific reasoning competencies of pre-service science teachers: Translating a German multiple-choice instrument into English and Spanish. *Int. J. Sci. Educ.* **2020**, *42*, 2819–2841. [CrossRef]
30. Mathesius, S.; Krell, M.; Upmeyer zu Belzen, A.; Krüger, D. Überprüfung eines Tests zum wissenschaftlichen Denken unter Berücksichtigung des Validitätskriteriums relations-to-other-variables. *Z. Päd.* **2019**, *65*, 492–511. [CrossRef]
31. Krüger, D.; Hartmann, S.; Nordmeier, V.; Upmeyer zu Belzen, A. Measuring Scientific Reasoning Competencies. In *Student Learning in German Higher Education: Innovative Measurement Approaches and Research Results*, 1st ed.; Zlatkin-Troitschanskaia, O., Pant, H.A., Toepper, M., Lautenbach, C., Eds.; Springer Fachmedien Wiesbaden: Wiesbaden, Germany, 2020; pp. 261–280. ISBN 9783658278854.
32. Opitz, A.; Heene, M.; Fischer, F. Measuring scientific reasoning—A review of test instruments. *Educ. Res. Eval.* **2017**, *23*, 78–101. [CrossRef]

33. Ding, L.; Wei, X.; Mollohan, K. Does Higher Education Improve Student Scientific Reasoning Skills? *Int. J. Sci. Math. Educ.* **2016**, *14*, 619–634. [CrossRef]
34. Krell, M.; Redman, C.; Mathesius, S.; Krüger, D.; van Driel, J. Assessing pre-service science teachers' scientific reasoning competencies. *Res. Sci. Educ.* **2020**, *50*, 2305–2329. [CrossRef]
35. Khan, S.; Krell, M. Scientific Reasoning Competencies: A Case of Preservice Teacher Education. *Can. J. Sci. Math. Technol. Educ.* **2019**, *19*, 446–464. [CrossRef]
36. Voss, T.; Kleickmann, T.; Kunter, M.; Hachfeld, A. Mathematics Teachers' Beliefs. In *Cognitive Activation in the Mathematics Classroom and Professional Competence of Teachers: Results from the COACTIV Project*; Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., Neubrand, M., Eds.; Springer: New York, NY, USA, 2013; ISBN 978-1-4614-5148-8.
37. Schoenfeld, A.H. Explorations of Students' Mathematical Beliefs and Behavior. *J. Res. Math. Educ.* **1989**, *20*, 338–355. [CrossRef]
38. Nouri, N.; Saberi, M.; McComas, W.F.; Mohammadi, M. Proposed Teacher Competencies to Support Effective Nature of Science Instruction: A Meta-Synthesis of the Literature. *J. Sci. Teacher Educ.* **2021**, *32*, 601–624. [CrossRef]
39. Kampurakis, K. The “general aspects” conceptualization as a pragmatic and effective means to introducing students to nature of science. *J. Res. Sci. Teach.* **2016**, *53*, 667–682. [CrossRef]
40. Neumann, I.; Kremer, K. Nature of Science und epistemologische Überzeugungen: Ähnlichkeiten und Unterschiede. *Z. Didakt. Naturwiss.* **2013**, *19*, 209–232.
41. Lederman, N.G.; Abd-El-Khalick, F.; Bell, R.L.; Schwartz, R.S. Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. *J. Res. Sci. Teach.* **2002**, *39*, 497–521. [CrossRef]
42. Liang, L.L.; Chen, S.; Chen, X.; Kaya, O.N.; Adams, A.D.; Macklin, M.; Ebenezer, J. Assessing preservice elementary teachers' views on the nature of scientific knowledge: A dual-response instrument: A Dual-Response instrument. *Asia-Pac. Forum Sci. Learn. Teach.* **2008**, *9*, 1–20.
43. Liang, L.L.; Chen, S.; Chen, X.; Kaya, O.N.; Adams, A.D.; Macklin, M.; Ebenezer, J. Preservice teachers' views about nature of scientific knowledge development: An international collaborative study. *Int. J. Sci. Math. Educ.* **2009**, *7*, 987–1012. [CrossRef]
44. Neumann, I.; Neumann, K.; Nehm, R. Evaluating Instrument Quality in Science Education: Rasch-based analyses of a Nature of Science test. *Int. J. Sci. Educ.* **2011**, *33*, 1373–1405. [CrossRef]
45. Neumann, K.; Härtig, H.; Harms, U.; Parchmann, I. Science teacher preparation in Germany. In *Model Science Teacher Preparation Programs: An International Comparison of What Works Best*; Pedersen, J., Isozaki, T., Hirano, T., Eds.; Information Age Publishing: Charlotte, NC, USA, 2017; pp. 29–52.
46. Kunz, H. Professionswissen von Lehrkräften der Naturwissenschaften im Kompetenzbereich Erkenntnisgewinnung [Professional Knowledge of Science Teachers in the Competence Area of Scientific Inquiry]. Ph.D. Thesis, University of Kassel, Kassel, Germany, 2012.
47. Duschl, R.A.; Grandy, R. Two Views about Explicitly Teaching Nature of Science. *Sci. Educ.* **2013**, *22*, 2109–2139. [CrossRef]
48. Khishfe, R.; Abd-El-Khalick, F. Influence of explicit and reflective versus implicit inquiry-oriented instruction on sixth graders' views of nature of science. *J. Res. Sci. Teach.* **2002**, *39*, 551–578. [CrossRef]
49. Bruckermann, T.; Aschermann, E.; Bresges, A.; Schlüter, K. Metacognitive and multimedia support of experiments in inquiry learning for science teacher preparation. *Int. J. Sci. Educ.* **2017**, *39*, 701–722. [CrossRef]
50. Krell, M.; Koska, J.; Penning, F.; Krüger, D. Fostering pre-service teachers' views about nature of science: Evaluation of a new STEM curriculum. *Res. Sci. Tech. Educ.* **2015**, *33*, 344–365. [CrossRef]
51. McDonald, C.V. The influence of explicit nature of science and argumentation instruction on preservice primary teachers' views of nature of science. *J. Res. Sci. Teach.* **2010**, *47*, 1137–1164. [CrossRef]
52. Mesci, G.; Schwartz, R.S. Changing Preservice Science Teachers' Views of Nature of Science: Why Some Conceptions May be More Easily Altered than Others. *Res. Sci. Educ.* **2017**, *47*, 329–351. [CrossRef]
53. Dogan, N.; Abd-El-Khalick, F. Turkish grade 10 students' and science teachers' conceptions of nature of science: A national study. *J. Res. Sci. Teach.* **2008**, *45*, 1083–1112. [CrossRef]
54. Nehring, A. Naïve and informed views on the nature of scientific inquiry in large-scale assessments: Two sides of the same coin or different currencies? *J. Res. Sci. Teach.* **2019**, *57*, 510–535. [CrossRef]
55. Khishfe, R. Explicit Nature of Science and Argumentation Instruction in the Context of Socioscientific Issues: An effect on student learning and transfer. *Int. J. Sci. Educ.* **2014**, *36*, 974–1016. [CrossRef]
56. Koenig, K.; Schen, M.; Bao, L. Explicitly Targeting Pre-Service Teacher Scientific Reasoning Abilities and Understanding of Nature of Science through an Introductory Science Course. *Sci. Educ.* **2012**, *21*, 1–9.
57. R Core Team. *R: A Language for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
58. Revelle, W. Psych: Procedures for Psychological, Psychometric, and Personality Research. 2021. Available online: <https://cran.r-project.org/web/packages/psych/index.html> (accessed on 1 August 2021).
59. Muthén, L.K.; Muthén, B.O. *Mplus*; Chapman and Hall/CRC: Los Angeles, CA, USA, 2021; Available online: https://www.statmodel.com/download/usersguide/MplusUserGuideVer_7.pdf (accessed on 1 August 2021).
60. Warm, T.A. Weighted likelihood estimation of ability in item response theory. *Psychometrika* **1989**, *54*, 427–450. [CrossRef]
61. Robitzsch, A.; Kiefer, T.; Wu, M. TAM: Test Analysis Modules. 2020. Available online: <https://cran.r-project.org/web/packages/TAM/index.html> (accessed on 1 August 2021).
62. Gajęcki, A.; Burzykowski, T. *Linear Mixed-Effects Models Using R*; Springer: New York, NY, USA, 2013; ISBN 978-1-4614-3899-1.

63. Nakagawa, S.; Schielzeth, H. A general and simple method for obtaining R² from generalized linear mixed-effects models. *Methods Ecol. Evol.* **2013**, *4*, 133–142. [CrossRef]
64. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Soft.* **2015**, *67*, 1–48. [CrossRef]
65. Rosenthal, R. *Meta-Analytic Procedures for Social Research*; SAGE Publications, Inc.: Thousand Oaks, CA, USA, 1991; ISBN 9780803942462.
66. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Lawrence Erlbaum: New York, NY, USA, 1988; ISBN 9781134742707.
67. Pinheiro, J.; Bates, D.; R-core. nlme: Linear and Nonlinear Mixed Effects Models. 2021. Available online: <https://cran.r-project.org/web/packages/nlme/index.html> (accessed on 1 August 2021).
68. Barton, K. MuMIn: Multi-Model Inference. 2020. Available online: <https://r-forge.r-project.org/projects/mumin/> (accessed on 1 August 2021).
69. Lenth, R.V. Emmeans: Estimated Marginal Means, aka Least-Squares Means. 2021. Available online: <https://cran.r-project.org/web/packages/emmeans/index.html> (accessed on 1 August 2021).
70. Ben-Shachar, M.S.; Makowski, D.; Lüdtke, D. Effectsize: Indices of Effect Size and Standardized Parameters. *J. Open Source Softw.* **2021**, *5*. [CrossRef]
71. Rosseel, Y.; Jorgensen, T.D.; Rockwood, N. Lavaan: Latent Variable Analysis. 2021. Available online: <https://cran.r-project.org/web/packages/lavaan/lavaan.pdf> (accessed on 1 August 2021).
72. Herman, B.C.; Clough, M.P. Teachers' longitudinal NOS understanding after having completed a science teacher education program. *Int. J. Sci. Math. Educ.* **2016**, *14*, 207–227. [CrossRef]
73. Kunina-Habenicht, O.; Schulze-Stocker, F.; Kunter, M.; Baumert, J.; Leutner, D.; Förster, D.; Lohse-Bossenz, H.; Terhart, E. Die Bedeutung der Lerngelegenheiten im Lehramtsstudium und deren individuelle Nutzung für den Aufbau des bildungswissenschaftlichen Wissens [The significance of learning opportunities in teacher training courses and their individual use for the development of educational-scientific knowledge]. *Z. Päd.* **2013**, *59*, 1–23.
74. Mathesius, S.; Hartmann, S.; Upmeier zu Belzen, A.; Krüger, D. Scientific Reasoning as an Aspect of Pre-service Biology Teacher Education. Assessing competencies using a paper-pencil test. In *The Future of Biology Education Research*; Tal, T., Yarden, A., Eds.; Technion—Israel Institute of Technology: Haifa, Israel, 2016; pp. 93–110.
75. Kuhn, T.S. *The Structure of Scientific Revolutions*, 2nd ed.; The University of Chicago Press: Chicago, IL, USA, 1970; ISBN 0-226-45803-2.
76. Gyllenpalm, J.; Wickman, P.-O. The Uses of the Term Hypothesis and the Inquiry Emphasis Conflation in Science Teacher Education. *Int. J. Sci. Educ.* **2011**, *33*, 1993–2015. [CrossRef]
77. Edgerly, H.S.; Kruse, J.W.; Wilcox, J.L. Quantitatively Investigating Inservice Elementary Teachers' Nature of Science Views. *Res. Sci. Educ.* **2021**, 1–14. [CrossRef]
78. Mathesius, S.; Upmeier zu Belzen, A.; Krüger, D. Eyetracking als Methode zur Untersuchung von Multiple-Choice-Aufgaben zum wissenschaftlichen Denken [Eye tracking as a method for studying multiple-choice scientific reasoning tasks]. In *Lehr- und Lernforschung in der Biologiedidaktik*; Hammann, M., Lindner, M., Eds.; StudienVerlag: Innsbruck, Austria, 2018; pp. 225–244.
79. Marton, F.; Tsui, A.B.M. *Classroom Discourse and the Space of Learning*; Lawrence Erlbaum: Mahwah, NJ, USA, 2004; ISBN 0805840087.
80. Hodson, D. Learning Science, Learning about Science, Doing Science: Different goals demand different learning methods. *Int. J. Sci. Educ.* **2014**, *36*, 2534–2553. [CrossRef]

Article

Individual Differences in Children's Scientific Reasoning

Erika Schlatter ^{1,*}, Ard W. Lazonder ¹, Inge Molenaar ¹ and Noortje Janssen ²

¹ Behavioural Science Institute, Radboud University, 6525 XZ Nijmegen, The Netherlands; ard.lazonder@ru.nl (A.W.L.); inge.molenaar@ru.nl (I.M.)

² Department of Instructional Technology, University of Twente, 7522 NB Enschede, The Netherlands; noortje.janssen@ru.nl

* Correspondence: erika.schlatter@ru.nl

Abstract: Scientific reasoning is an important skill that encompasses hypothesizing, experimenting, inferencing, evaluating data and drawing conclusions. Previous research found consistent inter- and intra-individual differences in children's ability to perform these component skills, which are still largely unaccounted for. This study examined these differences and the role of three predictors: reading comprehension, numerical ability and problem-solving skills. A sample of 160 upper-primary schoolchildren completed a practical scientific reasoning task that gauged their command of the five component skills and did not require them to read. In addition, children took standardized tests of reading comprehension and numerical ability and completed the Tower of Hanoi task to measure their problem-solving skills. As expected, children differed substantially from one another. Generally, scores were highest for experimenting, lowest for evaluating data and drawing conclusions and intermediate for hypothesizing and inferencing. Reading comprehension was the only predictor that explained individual variation in scientific reasoning as a whole and in all component skills except hypothesizing. These results suggest that researchers and science teachers should take differences between children and across component skills into account. Moreover, even though reading comprehension is considered a robust predictor of scientific reasoning, it does not account for the variation in all component skills.

Keywords: scientific reasoning; primary education; individual differences



Citation: Schlatter, E.; Lazonder, A.W.; Molenaar, I.; Janssen, N. Individual Differences in Children's Scientific Reasoning. *Educ. Sci.* **2021**, *11*, 471. <https://doi.org/10.3390/educsci11090471>

Academic Editors: Moritz Krell, Andreas Vorholzer and Andreas Nehring

Received: 29 June 2021
Accepted: 24 August 2021
Published: 27 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Science education is an important part of the curriculum in many countries [1,2]. Starting in primary school, children learn about the underlying principles and causal relationships of science domains as well as the processes through which this knowledge is created. This process of intentional knowledge-seeking is known as scientific reasoning [3,4] and is important for children because it prepares them for a society where science and the outcomes of scientific research are embedded in the culture [5]. In a school setting, scientific reasoning skills are particularly important for successful inquiry learning: 'minds-on' scientific reasoning skills [6] are instrumental to achieving meaningful outcomes from a 'hands-on' inquiry.

Scientific reasoning consists of multiple component skills, namely, hypothesizing, experimenting and evaluating evidence, the latter of which can be further divided into inferencing, evaluating data and drawing conclusions [7,8]. These component skills emerge at a different age, tend to develop at a different pace and are known to vary greatly between same-age children (e.g., [9]). However, most existing research either treats scientific reasoning as a unitary construct or looks at one specific component of scientific reasoning—most often experimenting [10]. Therefore, the inter- and intra-individual differences are not yet well understood, and to this date, few guidelines exist for addressing these differences in primary science classrooms.

An important challenge in understanding individual differences in scientific reasoning is the valid measurement of its component skills. Even though scientific reasoning is

often taught in hands-on settings, it is mostly measured by paper-and-pencil tests. As performance-based testing circumvents many of the problems typically associated with written tests (see, for an overview, Harlen [11]), it might shed new light on the development of scientific reasoning in children. Using one such performance-based test, the current study set out to advance our insight into children's proficiency in different component skills of scientific reasoning, when applied in a practical, coherent inquiry setting in order to ultimately aid the development of teaching materials for various groups of learners in primary education.

1.1. Variation in Scientific Reasoning

As mentioned above, scientific reasoning comprises the skills of hypothesizing (the articulation of ideas about possible outcomes of an investigation), experimenting (the skills to design and perform experiments to test these hypotheses) and evaluating evidence (i.e., drawing valid conclusions). Evidence evaluation, in turn, involves inferencing (making a verbal interpretation of the gathered data), evaluating data (assessing measurement quality, for instance, to decide whether there are enough data to base a conclusion on), and drawing conclusions (using this information to make causal statements to answer the research question).

This multidimensionality is confirmed by psychometric models [12] and studies investigating one or more component skills point to substantial variation. Experimenting, for example, is relatively easy for children to learn: most pre-schoolers are capable of some systematic testing [13,14], and older children can be taught this skill successfully by both direct instruction [15,16] and guided inquiry [17,18]. Hypothesizing is more difficult for young children to learn [9,13,19], whilst evidence evaluation is the most difficult skill for them to acquire [4] and is also experienced as such [20].

Most of the studies on which this tentative order of difficulty is based examined a single skill at a single time point. A positive exception is the study by Piekny and Maehler [9], who inferred the age at which children learn hypothesizing, experimenting and evidence evaluation from cross-sectional data collected with children from kindergarten to grade 5 and found a similar build-up as described above. Still, this study used different types of tasks for the different component skills rather than one task that encompassed all component skills. Thus, the relative difficulty of the component skills of scientific reasoning is not fully understood yet.

Other studies indicate that not all children develop scientific reasoning proficiency at the same pace. In a large-scale cross-sectional study using written tests in grades 4–6, Koerber, Mayer, Osterhaus, Schwippert and Sodian [21] distinguished between naïve, intermediate and advanced conceptions of scientific reasoning and found that, although older children more often had advanced conceptions and less often naïve conceptions than younger children, all proficiency levels were present at all participating grade levels. The results of Piekny and Maehler [9] further suggest that this variation increases with age. For example, both the means and standard deviations of 'hypothesizing' were low in kindergarten but increased from grade 1 onward. This finding indicates that, although children's hypothesizing skills grow, the inter-individual variation increases accordingly. Thus, although children improve in scientific reasoning over the years, not all children improve equally or at the same time as their peers. Acknowledging and understanding these differences is vital for good science education.

To conclude, the component skills of scientific reasoning improve considerably during the primary school years [9,21], albeit with substantial variation. As not all subskills emerge at the same point in time and not all children develop their scientific reasoning proficiency at the same pace, the teaching of scientific reasoning in primary education is a challenging task. A profound understanding of how the component scientific reasoning skills develop can help teachers make scientific reasoning accessible for all children.

1.2. Explaining Variation in Scientific Reasoning

Although differences in the development of scientific reasoning are known to exist, the roots of the differences between children as well as differences in developmental patterns within children (i.e., differences across skills) are less clear. Children's cognitive characteristics account for part of the variation in scientific reasoning proficiency. Previous research provides evidence that reading comprehension, numerical ability and problem-solving skills contribute to scientific reasoning.

Reading comprehension most consistently explains children's overall scientific reasoning performance on written tests [21,22] as well as their ability to set up unconfounded experiments using the Control-of-Variables Strategy [23,24]. Van de Sande, Kleemans, Verhoeven and Segers [25] found that reading comprehension explained the variance in all component scientific reasoning skills, albeit not to the same extent: effect sizes ranged from medium ($r = 0.30$) for experimentation and drawing conclusions to large ($r = 0.47$) for hypothesis validation. Why reading comprehension is such a strong predictor is not entirely clear. Possibly, reasoning ability transcends the domains of reading and science [24,25], or a general understanding of the language of science is important for science learning [26]. However, it is also possible that the influence of reading comprehension is a consequence of test item format: most of the studies cited above used written tests that likely call upon children's reading skills, even though questions were sometimes read out loud. In light of these findings, reading comprehension can be considered an important predictor of scientific reasoning, but because past research heavily relied on the use of paper and pencil tests, further scrutiny of its role is warranted.

Numerical ability is often named as a prerequisite for scientific reasoning by national curriculum agencies [27,28] as well as scientists [29]—likely because scientific reasoning, in particular the evidence evaluation skills, involves reasoning about numerical data [9,22,30,31]. Yet, empirical evidence for this relation is scarce. Early work by Bullock and Ziegler [32] demonstrated that numerical intelligence predicts the growth of experimentation skills in primary schoolchildren, explaining almost 35 percent of the variance in a quadratic growth model. More recent studies found significant correlations between numerical ability and scientific reasoning [10,33]. However, as the latter studies treated scientific reasoning as a unitary construct, it is yet unclear whether numerical ability also predicts children's scientific reasoning, and if so, if it predicts all component skills of scientific reasoning to the same extent.

Children's problem-solving skill is another possible predictor of scientific reasoning. Klahr and Dunbar [34] characterized scientific reasoning as a process of rule induction, which inherently involves problem-solving. One could even argue that scientific reasoning is a form of problem-solving in itself: the problem is a need for specific knowledge, which is resolved through a systematic process of knowledge-seeking. Furthermore, as with the previous predictors, it seems plausible that problem-solving calls upon a person's reasoning skills and therefore predicts scientific reasoning. Although upper-primary schoolchildren are still incapable of formal abstract reasoning, they can solve problems that involve reasoning with concrete objects such as the nine-dots problem and the Tower of Hanoi [35]. Recent research supports these ideas: Mayer et al. [22] found that problem-solving predicted a substantial portion of the variance in children's scientific reasoning. Van de Sande et al. [25] further showed that this effect does not apply to all subskills: hypothesis validation and experimenting depended on problem-solving, whereas generating conclusions did not. As such, problem-solving may explain some but not all component scientific reasoning skills, and the extent to which the different component skills are predicted is yet unclear.

1.3. Research Questions and Hypotheses

Although the cited literature points to notable differences in children's scientific reasoning, most studies either addressed scientific reasoning as a single, albeit multifaceted construct or examined one of its subskills in isolation. Furthermore, most extant research

has been conducted using written tests. These instruments neither resemble the learning context nor scientific practice and therefore may not accurately gauge children's true ability in scientific reasoning [36]. Moreover, written tests of scientific reasoning can confound with reading comprehension, as children with better reading comprehension might perform better on such tests because the test itself involves reading. In order to extend our understanding of the relations between scientific reasoning and the cognitive characteristics discussed above, the subskills should be studied in tandem, preferably in an authentic whole-task setting that does not require children to read.

This study, therefore, aimed to identify and explain differences in children's ability to reason scientifically by means of a performance-based task so as to maximize authenticity and minimize the influence of reading skills. A sample of 160 upper-primary schoolchildren performed this task to gauge their proficiency in five scientific reasoning skills: hypothesizing, experimenting, making inferences, evaluating data and drawing conclusions. Performance differences were related to reading comprehension, numerical ability and problem-solving skills in order to answer the following research questions:

1. What amount of variation can be found in children's scientific reasoning?
2. To what extent is this variation explained by reading comprehension, numerical ability and problem-solving skills?

Based on previous research using written tests, it was expected that children would differ considerably in their overall scientific reasoning proficiency. Differences across the five subskills were also predicted to occur. Specifically, children were expected to be most proficient in experimentation, less proficient in hypothesizing and least proficient in the three evidence evaluation skills (inferencing, evaluating data and drawing conclusions). Reading comprehension, numerical ability and problem-solving skills were expected to explain a unique portion of the variance in scientific reasoning. Considering the alleged differences across subskills, these characteristics were expected to have differential effects.

2. Materials and Methods

2.1. Participants

A sample of 166 children attending the two highest grades of a primary school in a suburban area of the Netherlands participated in this study. Ages ranged from 8 years 11 months to 12 years 8 months. About 80% of the parents held a degree from a research university or university of applied sciences, and almost all children had at least one parent who was born in the Netherlands. Complete data were obtained for 160 of the 166 participating children (54% boys, $M_{\text{age}} = 11$ years 0 months, $SD = 9$ months); 84 of these children were in grade 5 (52% boys, $M_{\text{age}} = 10$ years 5 months, $SD = 7$ months) and 76 of them in grade 6 (55% boys, $M_{\text{age}} = 11$ years 7 months, $SD = 6$ months).

The school participated in a large-scale longitudinal research project that was approved by the ethics committee of the Faculty of Behavioural, Management and Social Sciences of the University of Twente. All participating children had passive parental consent, meaning that parents were informed and did not object to their child's participation in the study. The findings reported here were gathered during the third wave of data collection, which means that the sample was familiar with most tests. The school's science curriculum contained five annual hands-on science projects which enabled children to practice their scientific reasoning.

2.2. Materials

2.2.1. Scientific Reasoning Task

Children's scientific reasoning skills were gauged during a 20 min performance-based scientific reasoning task under supervision of a test administrator [19]. The task contained 15 questions and assignments (hereafter referred to as 'items'), 3 for each component scientific reasoning skill, which were organized in four inquiry cycles of increasing difficulty (for example, see Table 1). The task was administered orally in order to minimize the effects of reading and writing ability, and handouts were used to ensure uniformity in the data

children used to make inferences, evaluate data and draw conclusions. Children's answers and actions were registered by the test administrator for later scoring. Each of the items was worth one point, and a child could thus earn a maximum of three points per subskill. Total test scores could range from 0 to 15 points. The Cohen's κ inter-rater agreement of the answer scoring was 0.84.

Table 1. Example inquiry cycle.

Subskill ¹	Question
Experimenting	Can you figure out if it matters whether the surface is hard or soft? So you can be sure whether the ball without the mat bounces more, less or as much as with the extra mat.
Inferencing	In this box, the outcome was 'tick-tick', and in this box, the outcome was 'tick-tick-tick-tick'. Can you explain what the outcome of the experiment was?
Drawing conclusions	Can you be sure that balls <i>always</i> bounce more often on a hard surface?
Hypothesizing	The student who is next will also be completing this experiment. Imagine that student asks you to predict what the outcomes of their experiment will be. What would you say?

¹ This is an example of the first inquiry cycle of the bouncing balls version of the test. In other versions, only the variables would be different. Evaluating data were assessed in subsequent research cycles.

Three versions of this task were available, which differed exclusively with regard to the topic of investigation. In the *rolling balls* version, adapted from Chen and Klahr [16], children interacted with two inclined planes to find out how four dichotomous input variables (slope, starting point, surface and mass of the ball) influenced the distance balls travel after leaving a ramp. In the *bouncing ball* version, children investigated how four dichotomous variables (starting height, surface, mass of the ball and whether the ball was solid) affected the number of times a ball would bounce; the *cars* version had children set four features of rubber-band-powered toy cars (size of back wheels, axle size, diameter of the rubber band and tightness of the winding of the rubber band) in order to examine how far a car drives.

Children were assigned to the version they had not received in previous waves of data collection, and scores did not differ significantly between the three versions, $F(2, 157) = 0.08$, $p = 0.925$. Furthermore, a validation study [19] showed no effects of prior domain knowledge on the performance of any of the versions. This study also demonstrated that the test scores conform to a two-parameter Item-Response theory model and have an acceptable expected a posteriori (EAP) reliability of 0.59. As the component skills were each assessed by only three items of increasing complexity, internal consistency of the subscales could not meaningfully be calculated.

2.2.2. Reading Comprehension Test

Reading comprehension was measured by a standardized progress evaluation measure developed by Cito, the Dutch national testing agency [37]. Different versions are available for different grades, and the test has a measurement accuracy between 0.87 and 0.89 [37]. In all versions of the test, children had to read different types of mostly pre-existing texts, such as short stories, newspaper articles, advertisements and instruction manuals. The test consisted of 55 multiple choice items that, for example, required children to fill in the blanks, explain what a particular line in the text means or choose an appropriate continuation of a story. As participants in the current study were drawn from different grades, the version corresponding to their grade level was administered. The One Parameter Logistic Model [38] was used to transform children's answers into a person proficiency score that can be meaningfully compared across grades.

2.2.3. Numerical Ability Test

Numerical ability was gauged by a standardized progress evaluation measure that required children to add, subtract, multiply or divide one- and two-digit numbers by heart [39]. The test consists of 200 items of increasing difficulty and is highly reliable ($\alpha = 0.97$). Children worked on the test for 5 min and obtained 1 point for each correct answer.

2.2.4. Problem-Solving Test

A digital version of the Tower of Hanoi (adapted from Welsh [40]) was developed to assess children's problem-solving skills. The test required children to solve as many problems as they could in 7 min. One point was awarded for each solved problem, and reliability was high ($\alpha = 0.85$). The 20 problems required children to move differently sized disks from their starting position to their target position on the rightmost peg. Three simple rules limited the possible moves children could make: only one disk could be moved at a time, the disk could only be moved to an adjacent peg and it could never be placed on top of a smaller disk. The starting position differed per problem in order to assure a gradual increase from a minimum of 3 moves to solve the puzzle at Problem 1 to a minimum 15 moves at Problem 19. The target solution for each of the problems was a three- or four-disk tower on the rightmost peg. In order to prevent trial-and-error and provide children with an opportunity for a fresh start if they had trouble solving a certain problem, each unsolved puzzle would be automatically reset after 20 moves were made. Manual reset was not possible. To ensure that children would not finish the task ahead of time, the final problem was a 5-disk, 31-move problem. In practice, none of the children reached this final problem.

2.2.5. Procedure

Children were tested in their regular classrooms. First, teachers administered the reading comprehension and numerical ability tests on a whole-class basis, using the guidelines provided by the test publishers. When standardized testing was completed, the researchers administered the problem-solving test and the scientific reasoning task. The problem-solving test was administered in small groups. After a short explanation, children worked on the test for 7 min. The scientific reasoning task was administered individually and lasted about 20 min per child.

2.2.6. Data Analysis

Data were analyzed using IBM SPSS 25. In order to answer the first research question, variation in scientific reasoning was explored using descriptive statistics; relations between the five scientific reasoning subskills were analyzed using Pearson correlations and a within-subject analysis of variance (ANOVA), controlled for grade and gender. The second research question, which sought to reveal what accounts for the observed differences in scientific reasoning, was answered by means of correlational analyses and multivariate multiple regression analysis.

Table 2 presents the descriptive statistics of children's test performance. Preliminary analyses of three predictor skills indicated that the sixth-graders outperformed the fifth-graders in reading comprehension, $F(1, 158) = 14.18, p < 0.001$, partial $\eta^2 = 0.08$, numerical ability, $F(1, 158) = 8.02, p = 0.005$, partial $\eta^2 = 0.05$ and problem-solving, $F(1, 158) = 4.35, p = 0.039$, partial $\eta^2 = 0.03$. The cross-grade differences in scientific reasoning were minor, and were tested for statistical significance in the main analysis reported below.

Table 2. Descriptive statistics of children’s test scores.

Test Scores	Grade 5		Grade 6		Entire Sample	
	M	SD	M	SD	M	SD
Scientific reasoning	8.01	2.23	7.99	2.24	8.00	2.23
Hypothesizing	2.15	0.86	2.01	0.82	1.77	0.88
Experimenting	1.54	0.63	1.50	0.64	2.09	0.84
Inferencing	1.19	0.63	1.37	0.73	1.52	0.63
Evaluating data	1.32	0.95	1.38	0.80	1.28	0.68
Drawing conclusions	1.81	0.86	1.72	0.90	1.35	0.88
Reading comprehension	52.40	12.58	61.91	18.98	56.92	16.59
Numerical ability	84.42	19.84	95.26	28.23	89.57	24.72
Problem-solving	11.39	2.92	12.33	2.74	2.87	0.17

3. Results

In order to determine the extent to which scientific reasoning ability differs between children, the means and standard deviations of children’s test scores were examined. Overall test scores ranged from 2 to 13 points with an average of 8.00 ($SD = 2.23$). Scores on the subskills ranged from 0 to 3 except for inferencing, where the minimum score was 1 point. Means and standard deviations confirmed this differential ability and warranted further exploration as to what could explain this difference in scientific reasoning proficiency.

The mean scores in Table 2 point to variation in proficiency on the different subskills: on average, children appeared to be most proficient in experimenting and least proficient in evaluating data and drawing conclusions, while hypothesizing and inferencing held the middle ranks. A within-subject ANOVA, controlling for gender and grade, was conducted to test whether these differences were statistically significant. Multivariate results revealed an overall effect of subskill (Pillai’s trace = 0.46, $F(4, 153) = 32.50$, $p < 0.001$), but no interaction effects of subskill with gender (Pillai’s trace = 0.02, $F(4, 153) = 0.60$, $p = 0.665$), and grade (Pillai’s trace = 0.03, $F(4, 153) = 1.23$, $p = 0.300$). The differences between subskills were further explored in univariate analyses. Scores on experimenting were significantly higher than scores on all other subskills ($p < 0.01$). Scores on hypothesizing were significantly higher than scores on inferencing, evaluating data and drawing conclusions ($p < 0.05$). Scores on inferencing were significantly higher than scores on evaluating data ($p < 0.01$), but not scores on drawing conclusions ($p = 0.214$). Drawing conclusions and evaluating data, the two subskills with the lowest scores, were not significantly different from one another ($p = 0.993$).

Having established that there is variation in the extent to which children master the five scientific reasoning subskills, the next set of analyses sought to explain these differences from children’s reading comprehension, numerical ability and problem-solving skills. As shown in Table 3, the total scientific reasoning score correlated with all three factors, albeit moderately. Correlations at the subskill level paint a mixed picture. Reading comprehension was associated with all subskills except hypothesizing, numerical ability only correlated with evaluating data and problem-solving did not correlate with any of the subskills.

Multivariate multiple regression was used to further scrutinize the relations between the three predictor variables and the five scientific reasoning subskills. Multivariate test results showed no main effect for the control variables gender, Pillai’s trace = 0.01, $F(5, 150) = 0.35$, $p = 0.882$, partial $\eta^2 = 0.01$, and grade, Pillai’s trace = 0.05, $F(5, 150) = 1.60$, $p = 0.164$, partial $\eta^2 = 0.51$. Regarding the explanatory variables, a significant contribution of reading comprehension on scientific reasoning was found, Pillai’s trace = 0.17, $F(5, 150) = 6.28$, $p < 0.001$, partial $\eta^2 = 0.17$. Neither numerical ability, Pillai’s trace = 0.02, $F(5, 150) = 0.57$, $p = 0.725$, partial $\eta^2 = 0.02$, nor problem-solving skills, Pillai’s trace = 0.02, $F(5, 150) = 0.61$, $p = 0.694$, partial $\eta^2 = 0.02$, explained scientific reasoning to a significant degree. The between-subject effects of reading comprehension in Table 4 showed that reading comprehension accounted for a significant proportion of the variance in experi-

menting, inferencing, evaluating data and drawing conclusions, but not in hypothesizing. The regression coefficients further indicate that experimenting was most influenced by reading comprehension. Of the significantly predicted subskills, inferencing was least influenced by reading comprehension. Thus, although reading comprehension remains an important explanatory factor, it did not explain all scientific reasoning subskills uniformly.

Table 3. Correlations for predictors and scientific reasoning subskills.

	1	2	3	4	5	6	7	8	9
1. Scientific reasoning (total score)	—								
2. Hypothesizing	0.61 **	—							
3. Experimenting	0.58 **	0.15	—						
4. Inferencing	0.53 **	0.18 *	0.11	—					
5. Evaluating data	0.46 **	0.14	0.14	0.06	—				
6. Drawing conclusions	0.63 **	0.17 *	0.16 *	0.28 **	0.08	—			
7. Reading comprehension	0.39 **	0.15	0.29 *	0.18 *	0.31 **	0.20 *	—		
8. Numerical ability	0.18 *	0.13	0.04	0.09	0.18 *	0.07	0.27 **	—	
9. Problem-solving	0.17 *	0.07	0.08	0.09	0.13	0.11	0.14	0.15	—

* $p < 0.05$, ** $p < 0.01$.

Table 4. Reading comprehension as explanatory factor of the scientific reasoning subskills.

Subskills	β	t	p	95% CI	Partial η^2
Hypothesizing	0.008	1.82	0.071	(0.00, 0.017)	0.021
Experimenting	0.017	4.09	<0.001	(0.009, 0.025)	0.098
Inferencing	0.007	2.10	0.037	(0.000, 0.013)	0.028
Evaluating data	0.011	3.23	0.001	(0.004, 0.018)	0.064
Drawing conclusions	0.011	2.43	0.016	(0.002, 0.020)	0.037

4. Discussion

This study aimed to identify and explain differences in children's ability to reason scientifically. To this end, a performance-based scientific reasoning task was administered, and measures of reading comprehension, numerical ability and problem-solving skills were collected in a sample of 160 upper-primary children. Their scientific reasoning scores varied considerably, which indicates that not all children are equally proficient in performing these skills. Observed differences within children further suggest that the five scientific reasoning skills are not equally difficult to perform. These intra-individual differences were partially explained by reading comprehension but not by numerical ability or problem-solving skills.

Results regarding the first research question confirm the existence of variation in children's scientific reasoning: the inter-individual spread in total scores was considerable, and marked intra-individual differences were found for some subskills. The hypothesized proficiency pattern was confirmed: children in our sample were most proficient in experimenting, less proficient in hypothesizing and least proficient in inferencing, evaluating data and drawing conclusions. This is particularly important because, as Koerber and Osterhaus [10] argued, previous research has studied these component skills separately, often through written tests [22,25]. The present study thus confirms the differences in subskill difficulty during a comprehensive performance-based scientific reasoning task and suggests that children's relative proficiency at the subskill level is stable across test modalities (cf. [6]).

Of particular interest is that the component scientific reasoning skills were consistently but moderately associated with total task scores. This result raises the question as to what accounts for the error variance in these correlations. Part of it could be due to the psychometric qualities of the scientific reasoning task. As mentioned in Section 2.2.1, each component skill was assessed by only three items, so a meaningful analysis of internal scale consistency was deemed impossible. In the absence of this information, the mag-

nitude of the correlations should be considered with some caution. A more substantive interpretation is that the proficiency pattern described above does not apply similarly to all children: some will develop the component skills in the indicated order, whereas others will show a deviating developmental trajectory. As a consequence, fine-grained measures of separate component skills, if reliably measured, give a more accurate impression of children's proficiency in scientific reasoning than global measures and should be the preferred approach when assessment serves diagnostic purposes, for instance, to inform the design of instruction.

The observed variation in scientific reasoning was independent of children's grade level. This equivalence of task performance might be due to the fact that our sample had few opportunities to practice their scientific reasoning skills—the school offered them only five inquiry projects per year, whereas the daily language and math classes lead to grade differences in reading comprehension and numerical ability. A related explanation is that scientific reasoning develops slowly in general and in the upper-primary grades in particular (e.g., [9]). Although most children at this age advance in scientific reasoning [19], the inter-individual variation is considerable and prevents the minor cross-grade growth differences from becoming statistically significant. Alternative research methods such as longitudinal designs and person-centered approaches to data analysis are more sensitive to capturing developmental growth and are increasingly being applied in scientific reasoning research [41].

Reading comprehension explained part of the variance in scientific reasoning. This result is consistent with hypotheses and complements previous research that administered written tests of scientific reasoning (e.g., [22,25,26]). Thus, why did reading comprehension predict scientific reasoning on a performance-based test that makes minimal demands on reading skills? One explanation is that scientific reasoning and reading comprehension both draw on general language comprehension processes, in particular when scientific reasoning is measured through interactive dialogue. Another interpretation could be that reading comprehension is a proxy of general intelligence or academic attainment, which, in turn, is associated with scientific reasoning (e.g., [42]). In addition, relations have been found between scientific reasoning and verbal reasoning [24], as well as nonverbal reasoning [25] and conditional sentence comprehension [43]. In line with these findings, language-centered scientific reasoning interventions have been proposed [25,43] and have been found to be effective [44].

Our results further show that reading comprehension does not explain all component scientific reasoning skills to the same extent, which underscores the importance of assessing the constituent skills separately rather than merging them in a single overarching construct. The most striking finding in this regard is that hypothesizing was not related to reading comprehension, even though one would intuitively expect verbal reasoning to be associated with this skill. Although it is not entirely clear why hypothesizing and reading comprehension were not related, a possible explanation may lie in what children need to reason about: their own ideas about the world (as in hypothesizing) as opposed to building a situation model from given information (as in reading [45] as well as in interpreting outcomes). In hypothesizing, misconceptions and naive beliefs may interfere with the reasoning process, whereas the chance of such 'illogical' thoughts could be less pronounced when reasoning with given information.

Numerical ability did not predict children's scientific reasoning. Although there were sound theoretical reasons to assume that numerical ability would predict scientific reasoning, empirical evidence on this relation is either scarce and relatively recent [10] or involved a different math strand [32]. Thus, while numerical ability as operationalized in this study does not explain individual differences in scientific reasoning, future research might examine whether this independence generalizes across tasks and settings. Future research could also investigate whether different math skills (e.g., number sense, measurement) contribute to performance on a scientific reasoning task.

Children's problem-solving skills did not predict scientific reasoning either, possibly because of task incongruence. Jonassen [46] argued that the ease with which a problem is solved relies on individual differences between problem solvers and problem characteristics. A scientific inquiry is an ill-defined problem that requires a problem solver to combine strategies and rules to come to an unknown solution, whereas the Tower of Hanoi is a well-defined problem with a constrained set of rules and a known solution. Thus, although the Tower of Hanoi does involve problem-solving, it may be insufficiently sensitive to distinguish weak from strong problem solvers. Beyond problem characteristics, the problem representation [46] might explain why Mayer et al. [22] found that the very similar Tower of London problem explained scientific reasoning. Mayer et al. [22] used a multiple-choice paper-and-pencil version of this problem in which all manipulations had to be completed mentally, thus making a relatively straightforward problem rather difficult to solve. As such, this test may not have identified all children who could solve a Tower of London problem, but only those who were sufficiently good at reasoning to complete the problem mentally. The current study, by contrast, used a less demanding task that allowed for real-time manipulation and was programmed to make invalid moves impossible. This difference in task demands might explain why the current study did not show a relation between problem-solving and scientific reasoning while previous research showed such a relation. As understanding what explains specific subskills is only a recent endeavor [10,25], more research is needed to understand which component skills can be explained as well as why differential effects are found.

4.1. Limitations

This study has some limitations, which include the homogenous sample in terms of parental background and education, with highly educated parents being overrepresented. As these parents are more likely to intellectually stimulate their children, for example, by taking them to science museums [47], this might have given the participants in the current study a certain advantage compared to children whose parents are less educated. The observed variation in scientific reasoning was nevertheless considerable and would probably have been even more diverse if a more heterogeneous sample had been used. Future research should therefore incorporate more diverse samples to find out whether the present conclusions generalize to more typical groups of upper-primary schoolchildren.

Another limitation lies in the task used to assess numerical ability. Because there was no precedent as to what type of math skills would predict scientific reasoning, a lean task that assessed basic numerical operations was chosen because it seemingly matched the type of operations children had to carry out during the scientific reasoning task (e.g., counting, direct comparisons). A further advantage of this task was that it did not make demands on reading skills, which is particularly important because previous studies did not allow for untangling of scientific reasoning and reading comprehension. However, although the current task resembled the types of *operations* children had to carry out during the scientific reasoning task, no *reasoning* was required. The absence of any significant results suggests that numerical ability may not be the most relevant math skill to predict scientific reasoning, and further research is needed to identify if and what math skills relate to scientific reasoning.

4.2. Implications

The current study confirms that scientific reasoning is a multifaceted construct. This is not only evident from differences in children's proficiency in the component skills but also from the asymmetry in the extent to which reading comprehension predicts these skills. How children of different proficiency levels learn scientific reasoning in a classroom setting and can be taught to reach their best potential is something that needs to be attended to in future research. Studying all scientific reasoning skills together is particularly important. Previous research has predominantly focused on a single skill, most often experimenting [48], which stands to reason because experimenting is such

a fundamental skill. At the same time, these focused investigations do not capture the complexity of scientific inquiry, the relative proficiency of children in the different subskills and the relations between these skills. Therefore, future research should focus more on scientific reasoning in authentic inquiry settings while still distinguishing subskills.

The absence of grade-level differences suggests that scientific reasoning develops slowly in the upper-primary years and implies that sustained practice is needed to boost this development. In preparing weekly or bi-weekly inquiry-based science lessons, teachers should attend to differences between children and among subskills. Most children will be able to perform the relatively easy skill of experimenting themselves with minimal guidance, whereas more teacher guidance is needed in generating hypotheses. Inferencing, evaluating data and drawing conclusions, which are the most difficult subskills, should initially be taken over by the teacher, who can demonstrate the skills to the class and gradually decrease their involvement as the lesson series progresses.

Results of the multiple regression analysis imply that teachers who start an inquiry-based curriculum can infer children's entry levels from their reading comprehension scores—children's basic numerical skills and ability to solve mind puzzles that resemble the Tower of Hanoi (e.g., tangrams, sudokus) should not be used for this purpose because both are poor predictors of scientific reasoning. The regression data also suggest that proficient readers need less guidance in scientific reasoning, so teachers can devote more attention to the average and poor readers in the class. Teachers should, of course, monitor the progress of all children and adjust the level of guidance just-in-time on an as-needed basis. A final practical suggestion concerns the scheduling of inquiry-based science classes. As these lessons are often taught by specialist teachers with part-time contracts, schools can opt for flexible scheduling and combine the fifth- and sixth-grade lessons because the proficiency levels in these classes are comparable. Alternatively, the same lessons can be delivered in both grades, perhaps with some minor adjustments in the amount of guidance, which will ease the teachers' burden in lesson preparation.

5. Conclusions

This study found substantial overall differences in children's scientific reasoning as well as marked differences at the subskill level. This variation was in part explained by children's reading comprehension but not their numerical ability and problem-solving skills. These results confirm the importance of treating scientific reasoning as a multifaceted skill. Both teachers and researchers should address scientific reasoning in an integrated setting where its component skills are distinguished but not studied or taught in isolation. As reading comprehension explains scientific reasoning in general and most of its constituent skills, science teachers should give more guidance to the poor readers in their classes, and researchers should administer performance-based assessments of scientific reasoning that make minimal demands on reading skills.

Author Contributions: Conceptualization, E.S., A.W.L. and I.M.; methodology, E.S., A.W.L. and I.M.; software, E.S.; formal analysis, E.S.; investigation, E.S. and N.J.; resources, E.S. and A.W.L.; data curation, E.S.; writing—original draft preparation, E.S.; writing—review and editing, A.W.L., I.M. and N.J.; supervision, A.W.L. and I.M.; project administration, E.S. and A.W.L.; funding acquisition, A.W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Netherlands Initiative for Education Research (NRO), grant number 405-15-546.



Institutional Review Board Statement: This study was approved by the ethics committee of the Faculty of Behavioural, Management and Social Sciences of the University of Twente, under number 15460.

Informed Consent Statement: All participating children had passive parental consent, meaning that parents were informed and did not object to their child's participation in the study.

25. Van de Sande, E.; Kleemans, M.; Verhoeven, L.; Segers, E. The linguistic nature of children's scientific reasoning. *Learn. Instr.* **2019**, *62*, 20–26. [CrossRef]
26. Snow, C.E. Academic language and the challenge of reading for learning about science. *Science* **2010**, *328*, 450–452. [CrossRef]
27. van Graft, M.; Tank, M.K.; Beker, T.; van der Laan, A. *Wetenschap en Technologie in Het Basis- en Speciaal Onderwijs: Richtinggevend Leerplankader Bij Het Leergebied Oriëntatie op Jezelf en de Wereld.*; SLO (nationaal expertisecentrum leerplanontwikkeling): Enschede, The Netherlands, 2018.
28. Wong, V. Authenticity, transition and mathematical competence: An exploration of the values and ideology underpinning an increase in the amount of mathematics in the science curriculum in England. *Int. J. Sci. Educ.* **2019**, *41*, 1805–1826. [CrossRef]
29. Schauble, L. In the eye of the beholder: Domain-general and domain-specific reasoning in science. In *Scientific Reasoning and Argumentation*; Fischer, F., Chinn, C.A., Engelmann, K., Osborne, J., Eds.; Routledge: New York, NY, USA, 2018; pp. 11–33.
30. Krummenauer, J.; Kuntze, S. Primary students' reasoning and argumentation based on statistical data. In Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education, Utrecht, The Netherlands, 6–10 February 2019.
31. Makar, K.; Bakker, A.; Ben-Zvi, D. The reasoning behind informal statistical inference. *Math. Think. Learn.* **2011**, *13*, 152–173. [CrossRef]
32. Bullock, M.; Ziegler, A. Scientific reasoning: Developmental and individual differences. In *Individual Development from 3 to 12: Findings from the Munich Longitudinal Study*; Cambridge University Press: Cambridge, UK, 1999; pp. 38–54.
33. Tajudin, N.M.; Chinnappan, M. Exploring relationship between scientific reasoning skills and mathematics problem solving. In Proceedings of the 38th Annual Conference of the Mathematics Education Research Group of Australasia: Mathematics Education in the Margins, Sunshine Coast, Australia, 28 June–2 July 2015; pp. 603–610.
34. Klahr, D.; Dunbar, K. Dual space search during scientific reasoning. *Cogn. Sci.* **1988**, *12*, 1–48. [CrossRef]
35. Piaget, J. Part I: Cognitive development in children: Piaget development and learning. *J. Res. Sci. Teach.* **2003**, *40*, 8–18. [CrossRef]
36. Shavelson, R.J.; Baxter, G.P.; Pine, J. Performance assessment in science. *Appl. Meas. Educ.* **1991**, *4*, 347–362. [CrossRef]
37. Weekers, A.; Groenen, I.; Kleintjes, F.; Feenstra, H. *Wetenschappelijke Verantwoording Papieren Toetsen Begrijpend Lezen Voor Groep 7 en 8 [Scientific Justification Pen-and-Paper Tests Reading Comprehension Grade 5 and 6]*; CITO: Arnhem, The Netherlands, 2011.
38. Verhelst, N.D.; Glas, C.A.W. The one parameter logistic model. In *Rasch Models: Foundations, Recent Developments and Applications*; Fischer, G.H., Molenaar, I.W., Eds.; Springer: New York, NY, USA, 1995; pp. 215–239. [CrossRef]
39. De Vos, T. *Schoolvaardigheidstoets Hoofdrekenen [Arithmetic Proficiency Test for Primary School]*; Boom Test Uitgevers: Amsterdam, The Netherlands, 2006.
40. Welsh, M.C. Rule-guided behavior and self-monitoring on the tower of hanoi disk-transfer task. *Cogn. Dev.* **1991**, *6*, 59–76. [CrossRef]
41. Hickendorff, M.; Edelsbrunner, P.A.; McMullen, J.; Schneider, M.; Trezise, K. Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learn. Individ. Differ.* **2018**, *66*, 4–15. [CrossRef]
42. Veenman, M.V.J.; Wilhelm, P.; Beishuizen, J.J. The relation between intellectual and metacognitive skills from a developmental perspective. *Learn. Instr.* **2004**, *14*, 89–109. [CrossRef]
43. Svirko, E.; Gabbott, E.; Badger, J.; Mellanby, J. Does acquisition of hypothetical conditional sentences contribute to understanding the principles of scientific enquiry? *Cogn. Dev.* **2019**, *51*, 46–57. [CrossRef]
44. Van der Graaf, J.; van de Sande, E.; Gijssels, M.; Segers, E. A combined approach to strengthen children's scientific thinking: Direct instruction on scientific reasoning and training of teacher's verbal support. *Int. J. Sci. Educ.* **2019**, *41*, 1119–1138. [CrossRef]
45. Swart, N.M.; Muijselaar, M.M.L.; Steenbeek-Planting, E.G.; Droop, M.; de Jong, P.F.; Verhoeven, L. Cognitive precursors of the developmental relation between lexical quality and reading comprehension in the intermediate elementary grades. *Learn. Individ. Differ.* **2017**, *59*, 43–54. [CrossRef]
46. Jonassen, D.H. Toward a design theory of problem solving. *Educ. Technol. Res. Dev.* **2000**, *48*, 63–85. [CrossRef]
47. Archer, L.; Dawson, E.; Seakins, A.; Wong, B. Disorientating, fun or meaningful? Disadvantaged families' experiences of a science museum visit. *Cult. Stud. Sci. Educ.* **2016**, *11*, 917–939. [CrossRef]
48. Rönnebeck, S.; Bernholt, S.; Ropohl, M. Searching for a common ground—A literature review of empirical research on scientific inquiry activities. *Stud. Sci. Educ.* **2016**, *52*, 161–197. [CrossRef]

Article

Reasoning on Controversial Science Issues in Science Education and Science Communication

Anna Beniermann *, Laurens Mecklenburg and Annette Upmeier zu Belzen 

Biology Education, Humboldt-Universität zu Berlin, 10099 Berlin, Germany;

laurens.mecklenburg@web.de (L.M.); annette.upmeier@biologie.hu-berlin.de (A.U.z.B.)

* Correspondence: anna.beniermann@hu-berlin.de

Abstract: The ability to make evidence-based decisions, and hence to reason on questions concerning scientific and societal aspects, is a crucial goal in science education and science communication. However, science denial poses a constant challenge for society and education. *Controversial science issues* (CSI) encompass scientific knowledge rejected by the public as well as *socioscientific issues*, i.e., societal issues grounded in science that are frequently applied to science education. Generating evidence-based justifications for claims is central in scientific and informal reasoning. This study aims to describe attitudes and their justifications within the argumentations of a random online sample ($N = 398$) when reasoning informally on selected CSI. Following a deductive-inductive approach and qualitative content analysis of written open-ended answers, we identified five types of justifications based on a fine-grained category system. The results suggest a topic-specificity of justifications referring to specific scientific data, while justifications appealing to authorities tend to be common across topics. Subjective, and therefore normative, justifications were slightly related to conspiracy ideation and a general rejection of the scientific consensus. The category system could be applied to other CSI topics to help clarify the relation between scientific and informal reasoning in science education and communication.

Keywords: argumentation; reasoning; justifications; socioscientific issues; societally denied science; controversial science issues; science communication; science education

Citation: Beniermann, A.; Mecklenburg, L.; Upmeier zu Belzen, A. Reasoning on Controversial Science Issues in Science Education and Science Communication. *Educ. Sci.* **2021**, *11*, 522. <https://doi.org/10.3390/educsci11090522>

Academic Editors: Moritz Krell, Andreas Vorholzer and Andreas Nehring

Received: 8 August 2021

Accepted: 3 September 2021

Published: 8 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The OECD Learning Compass 2030 [1] highlights the rapid changes confronting our society and, consequently, the importance of adaptive education in formal and informal learning environments. It emphasizes the need to think and act responsibly “towards collective well-being” [1] based on knowledge, attitudes, values, and skills (including reasoning and critical thinking) as a 21st-century goal [1]. In contrast to this goal, science denial poses a constant or even growing challenge for society [2] and science education [3]. Informed citizens should be able to make evidence-based decisions on questions concerning scientific and societal aspects, e.g., health and environmental issues [4].

The inevitable connections between science and society in science education are bundled under the term *socioscientific issues* (SSI), defined as “societal issues with conceptual or technological ties to science” [5]. SSI are scientific topics that are often discussed controversially by the public [6]. They are well-acknowledged as contexts for science learning [7,8], as the SSI approach integrates scientific, sociological, and ethical content to foster reasoning on complex questions [9]. For example, the current COVID-19 pandemic illustrates the rise in controversy between society and science and, moreover, in doubt about scientific findings [10].

While the scientific foundation of some SSI is mostly accepted by the public (e.g., knowledge about stem cells), controversy may arise with the ethical dilemmas of its application (e.g., stem cell research for medical purposes) [11]. Other SSI are based on societally controversial science that may even be rejected by parts of the public (e.g., anthropogenic causes

of climate change), while being quite undisputed among scientists [11]. These topics are referred to as *societally denied science* [11] or *controversial science issues* (CSI) [12]. Attitudes toward CSI, i.e., their rejection or acceptance, highly rely on individual norms and values that do not necessarily result from scientific reasoning [13].

As science and technology develop rapidly, opportunities to encounter a variety of SSI, on which decisions must be made, and CSI, on which attitudes must be formed, become more frequent. Fostering the ability to make informed decisions on such complex issues and problems using evaluation and reasoning is not only a crucial aspect of general scientific literacy [14,15] but also a central goal of science education [16,17] and science communication [18,19]. Scientific reasoning [20,21] and informal reasoning (i.e., everyday reasoning on ill-structured problems) [22] on SSI and CSI entail the evaluation and justification of claims.

Several researchers have pointed out that argumentation is a core competence for reasoning and scientific inquiry [23], as well as central to science education in general [24]. There are few studies on argumentation in science communication [25], but it is a potentially beneficial field to bridge science communication and science education [18]. Argumentation in terms of SSI and CSI involves an ethical dimension, so socioscientific argumentation is a distinct process from scientific argumentation [26]. Multiple studies have demonstrated that reasoning on SSI [27,28] improves the complexity and quality of students' arguments concerning both scientific and socioscientific issues and can improve students' argumentation skills [29] and critical scientific literacy [30].

Different approaches to assessing argumentation have been used in science education [31]. *Toulmin's Argument Pattern* (TAP) [32] in particular has been applied in various ways [33]. However, TAP is predominantly used to assess the quality of students' arguments [34] and focuses on an argument's structure [35]. To date, few studies have examined the content of arguments [27,35] or justifications [36].

Furthermore, most research on informal reasoning and argumentation in the context of SSI and CSI focuses on either school [27–29,33,35] or university students [16,36,37], but similar analytical approaches to argumentations used by the public [18] could provide insights into controversial debates on scientific issues in everyday life. This research aims to describe different kinds of justifications used when people reason informally on selected CSI based on their attitude, using a fine-grained category system.

2. Theoretical Background

2.1. Socioscientific Issues (SSI) and Controversial Science Issues (CSI)

“Controversial Science Issues are scientific topics that, by their very nature, create discussions, debates, and questions because students are intrigued by these issues, question them or even have significant doubts about them” [12] (p. 26). Often, the description of controversy in the relationship between science and society is left implicit in science communication [38,39] as well as science education [36]. Borgerding and Dagistan [11] differentiate between three categories of CSI: *active science*, *societally denied* and *societally accepted science*, and SSI (Figure 1).

Controversies within active science are located within the scientific community itself (actual scientific frontier debates) [11]. Societally denied science refers to a negative attitude (i.e., rejection) toward scientific knowledge among the public (i.e., “x is not true”). This scientific knowledge was nevertheless generated within the scientific community and a scientific consensus on it exists [11].

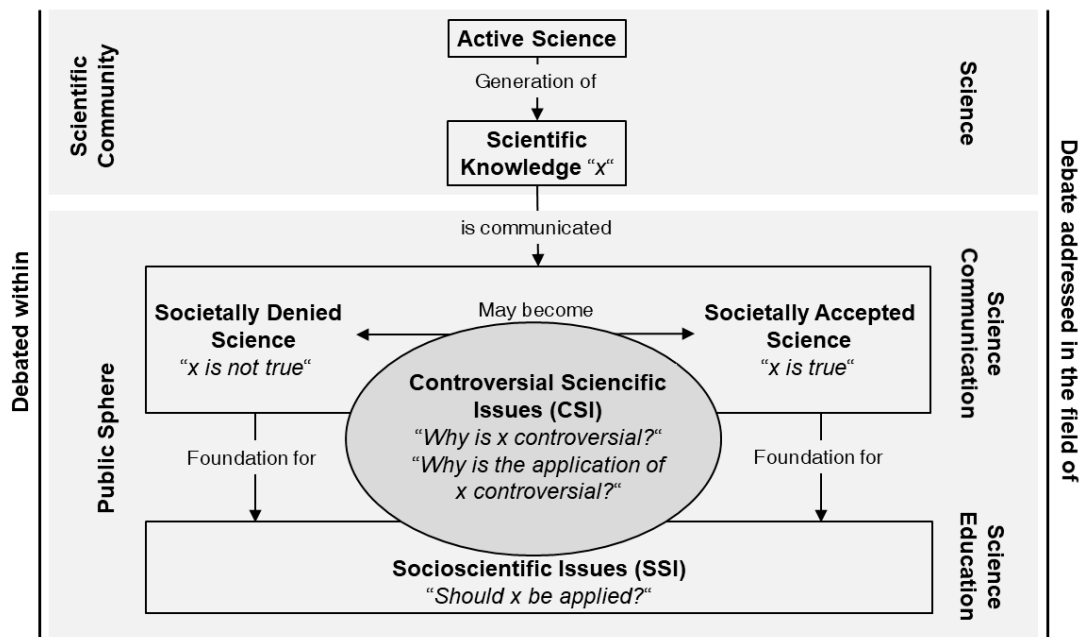


Figure 1. Relationship between controversial science issues, active science, societally denied and accepted science, and socioscientific issues (adapted from [11]).

Both societally denied science and societally accepted science can serve as a foundation for SSI [11]. For instance, knowledge about stem cell research counts as societally accepted science [11], even if the application of this research can be addressed as controversial when teaching SSI. SSI are highly relevant to society and are often discussed controversially in the context of science education [5]. Examples of SSI include stem cell research, environmental issues and their possible solutions, and the creation of genetically modified organisms [40], and therefore, the applications of scientific knowledge in these areas [11]. Normative questions, like "Is the application of this technology just?" are addressed in typical SSI, reflecting the fact that SSI cannot be resolved by science and scientific inquiry practices alone [11]. SSI have an ethical dimension concerning the relationship between science, technology, and society [41] as well as a complex societal dimension [5]. Problems in the context of SSI are open-ended, ill-structured, subject to multiple perspectives, and they lack clear solutions [37].

However, not all controversial scientific topics addressed in educational contexts fit this definition of SSI. Issues may be controversial and contested within the public sphere without being ill-structured and/or without lacking clear solutions. SSI are often described as inherently controversial [9,35] or as one kind of controversial issue [27,42]. Following the ideas that SSI are one kind of CSI and that a publicly contested issue is not necessarily denied by the public, we describe CSI as an umbrella term (see Figure 1), comprising different approaches of science communication and science education. When engaging in CSI, the question is not whether a certain issue is true or just but what the reasons for its controversy are.

Examples of CSI are evolution [43] and climate change [44], since parts of society doubt their theoretical scientific foundation, i.e., have a negative attitude toward them. Attitudes are conceptualized here as an affective assessment of an attitude object (e.g., evolution, climate change) [45]. Nevertheless, these topics do not lack a clear solution and are not ill-structured. For other CSI, such as vaccination [46] and GMOs [47], the controversy refers, at least in part, to the application of technology and touches the field of SSI.

Moreover, different factors influence attitudes toward CSI topics, hence the distinction between societally denied and societally accepted science. Most influencing factors that affect the rejection of scientific knowledge or applications are affective, such as emotions,

ideology, or worldview, and are referred to as the roots of attitudes [13]. Attitudes toward vaccination depend on risk perception, barriers, trust, calculation, and responsibility for society [46], while factors like religious belief [43], trust in science, and knowledge about the *nature of science* (NOS) influence attitudes toward evolution [48]. Climate change attitudes are influenced by political identity [49] and an individualistic worldview [50], and attitudes toward GMOs are affected by views about natural purity [51] as well as emotions and intuitions [52]. These different factors illustrate the issue-dependency and high heterogeneity of predictors of the controversiality of a topic [53,54]. However, some factors seem to be general predictors of the acceptance and rejection of scientific knowledge, like conspiracy ideation [55] and knowledge about NOS [56].

2.2. Informal and Scientific Reasoning

Engagement in SSI often involves argumentation and decision-making processes that require reasoning processes, i.e., processes of building and evaluating arguments [57]. For a long time, research on reasoning focused on formal reasoning about well-defined problems [58] and followed a “deduction paradigm” [59]. However, it has been demonstrated that human reasoning is prone to biases, and everyday reasoning is in most cases informal reasoning [58]. Both formal (scientific) and informal reasoning are processes of generating and assessing arguments [60]. While the problems addressed in scientific reasoning are often well-defined and the respective premises are explicit, problems in informal reasoning are ill-structured and the premises are not always stated [61]. Informal reasoning tasks often involve generating and evaluating positions on complex issues that lack clear solutions [5]. However, the coordination of theory and evidence [4,60], as well as generating evidence-based justifications [60], is central in informal and scientific reasoning: “Foundational abilities that lie at the heart of both types of reasoning are the ability to recognize the possible falsehood of a theory, and the identification of evidence capable of disconfirm” [60] (p. 74). These abilities align with the epistemic dimension of scientific reasoning as described by Osborne [21].

As SSI typically involve contentious and open-ended problems, their negotiation and resolution can be characterized by informal reasoning [5,61], which is especially suitable for processes like decision-making about actions for which supporting and opposing arguments exist [57]. The ability to informally reason on SSI has been described as a crucial component of scientific literacy [5] and a central goal of science education [62].

In addition to components of scientific reasoning [20,21], reasoning on SSI requires the integration of societal and ethical aspects, also referred to as moral reasoning [63,64]. Sadler, Barab, and Scott [8] proposed the construct of *socioscientific reasoning* (SSR) to assess the reasoning practices associated with SSI. While research on SSR highlights the integration of ethical components that require moral reasoning [42], reasoning on CSI is not necessarily a matter of moral reasoning but a matter of personal attitudes and knowledge. This is because the questions concerning CSI are neither open-ended nor unsolvable dilemmas [11]. Therefore, frameworks developed to assess SSR competencies [65], decision-making on SSI [66], and SSI attitudes [67] cannot be applied to CSI in which a clear scientific consensus concerning scientific knowledge and/or its application has been reached. Assessing how people reason concerning their attitude toward a CSI asks for different approaches, e.g., the identification of informal reasoning types [37,61,68]. While some research results suggest that reasoning is consistent across different topics [65,69], other studies describe a topic-specificity [70,71].

2.3. Argumentation Frameworks

Argumentation is the communicative part of reasoning [22] and is addressed more and more by science curricula around the world [33]. Argumentation in science is an essential skill, not only for scientists and science students but also for citizens, to enable them to make informed decisions on (socio-)scientific issues in everyday life [33].

Argumentation in general can be described as an interplay of constructing claims or explanations and the corresponding evidence [32,72] to justify something [73]. A fine-grained conceptualization of argumentation has been an ongoing challenge for researchers, and a variety of frameworks exist [31]. Aside from differences among these frameworks focusing either on content [28], structure [74], or the epistemological quality [75] of arguments, all of these frameworks rely on Toulmin's Argument Pattern (TAP) [32]. The TAP builds a general structure of arguments (Figure 2) and a foundation to assess them [33].

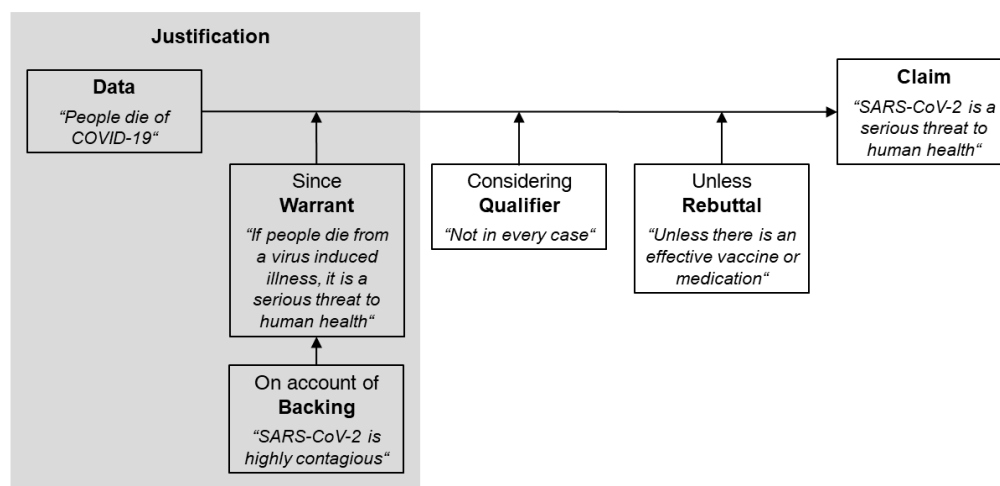


Figure 2. Toulmin's Argumentation Pattern (adapted from [32,33]) and its application to a complex argument concerning the CSI of SARS-CoV-2.

The *claim* of an argument is its conclusion. It is a statement of commitment [33] that every individual can agree or disagree with. The claim is based on several elements of the argument, with *data* representing the evidence for the claim being the central element. The data needs a *warrant* as a conclusive rule, turning the data into a relevant reason to support the claim. The warrant itself can furthermore be based on additional information called a *backing*. Because those three compartments form the justifying part of a persuasive argument [32] they are subsequently subsumed as the *justification* of a claim. The justification is opposed to the *rebuttal*, which contradicts it, and the *qualifier*, which describes the extent to which the justification allows valid conclusions.

As humans are easily capable of connecting statements in a logical way, the warrant is sometimes left implicit [76]. In the given example, the fact that people die from COVID-19 (i.e., data) can lead to the conclusion of SARS-CoV-2 posing a serious threat to human health without formulating the warrant (i.e., that the possible death forms a serious threat to human health). Equally, the data can be left implicit. Taking this into account, the articulation of a justification does not always include both data and warrant but sometimes appears as only one of the two components.

The TAP is often used as an analytical framework to evaluate argument quality [33]. When assessing arguments, an adapted version of TAP is often used. Qualifiers are often neglected to reduce the complexity [77–79]. The claim-evidence-reasoning approach is an established adaptation of the framework, in which, in addition to claim and evidence (i.e., data), warrant and backing are summarized as reasoning [79].

However, using TAP or its adapted forms as an analytical tool has also been criticized [33] due to the ambiguity of the arguments' elements [80] and the context-dependency of their interpretation [81]. In particular, differentiating between data and warrant, as well as warrant and backing, is difficult and depends on the context [77,82]. These challenges, as well as approaches that merge data and warrant [82], underpin the justification component (see Figure 2).

Several other studies that assessed arguments did not rely on TAP but analyzed arguments dichotomously by focusing on one claim supported by a ground, i.e., a reason [28,36]. Often, *subjective* and *objective* justifications are distinguished [36,83]. A comparable distinction was provided by Jafari and Meisert [27], who distinguished between normative and fact-based reasoning. Objective justifications are sometimes further divided into *evidential* and *deferential* justifications [83], with deferential justifications appealing to an authority [83]. Justifications were found to be heterogeneous within a person's argumentation and to differ among different CSI [36].

Additionally, several studies have indicated the relation between knowledge about NOS and argumentation skills [84,85]. Studies on argumentations in the field of SSI predominantly focus on argument quality based on the TAP or adapted forms. However, when it comes to the argument's contents and the types of justifications within arguments, few studies are available [27,86].

2.4. Research Questions

Several researchers have pointed out that instruction and conceptual knowledge of argumentation can foster the use of more complex [28] and more fact-based [27] arguments in the science classroom. There are still societal debates on scientific topics that are not disputed in the scientific sphere and do not lack clear solutions, and these topics count as CSI. Scientific knowledge, or its application that is partly rejected by the public, points to negative attitudes toward a topic. As roots of such attitudes are known to be mostly affective [13], this leads to the question of how people justify these attitudes.

An assessment of justifications within arguments on CSI in the public sphere is a necessary first step to identify overall tendencies and context-dependencies in justifications as one element of informal reasoning. In the long run, a resulting framework may help equip students with the necessary skills to participate in these public debates.

Our study addresses the following research questions:

1. To what extent can justifications identified in the field of CSI be grouped, with regard to theoretical criteria? (RQ1)
2. To what extent are justifications specific for certain CSI (topic-specific)? (RQ2)
3. How are acceptance and rejection of CSI related to the use of different justifications? (RQ3)
4. How does knowledge about NOS, religiousness, and conspiracy ideation relate with the use of different justifications? (RQ4)

3. Materials and Methods

3.1. Participants and Data Collection

We conducted an online survey in German, distributed via social networks to reach the public in an informal learning context. Postings included a short introduction to the topic and targeted communities interested in CSI, e.g., through a comment on videos on genetically modified food (GMF; YouTube), anti-vaccine groups (Facebook), and science communicators (Twitter). This random sampling was justified by the aim to reach out to a heterogeneous sample and collect a wide range of different justifications on different CSI. Data were collected within a two-month period in summer 2020.

In total $N = 398$ volunteers took part in the survey, of which $N = 265$ completed the questionnaire up to the last page. Participation in the survey was voluntary and during free time, which might explain the high dropout rate. It was possible to skip questions or leave the survey at any point. All open answers were analyzed, regardless of if the data set was complete. For closed questions, listwise deletion was applied. The age of participants ranged from 16 to 85 with an average of 41 years. Participants from all 16 German provinces took part, and the education level ranged from high school students and people who left school without a degree to post-doctoral researchers.

3.2. Instruments

The research design was adapted from Lobato and Zimmerman [36]. Their survey included four CSI topics (evolution, climate change, GMF, vaccination) and involved confronting participants with a statement (i.e., claim) reflecting the scientific consensus on each topic (Table 1). We added a fifth statement on SARS-CoV-2 to the survey, as the pandemic has led to the most substantive large-scale, open, and public discussion of epidemiology and science in recent history [87]. The statements used by Lobato and Zimmerman [36] were modified whenever they seemed to express epistemological considerations that could also serve as justifications, like “Evolution is the best explanation” or “Medical research has demonstrated” (Table 1).

Table 1. Statements reflecting scientific consensus on five CSI topics (modified based on [36]). Statements reflect the claim in TAP [32].

CSI Topic	Original Statement [36]	Adjusted Statement	Adjusted Statement (German)
Evolution	Biological evolution is the best explanation for explaining the varieties of species of life.	The variety of life forms and species is rooted in evolution.	Die Vielfalt an Lebensformen und Arten ist auf Evolution zurückzuführen.
Climate Change	The earth is experiencing a period of global climate change that human activity is contributing to.	The earth is experiencing a period of global climate change that human activity is largely contributing to.	Die Erde unterliegt einem klimatischen Wandel, zu dem der Mensch maßgeblich beiträgt.
Genetically modified foods (GMF)	Genetically modified foods [also known as GM or GMO foods] are largely safe for human consumption.	Genetically modified foods are largely safe for human consumption.	Genetisch veränderte Lebensmittel sind größtenteils sicher für den menschlichen Verzehr.
Vaccination	Medical research has demonstrated that childhood vaccinations are largely safe and effective.	Vaccinations are largely safe and effective.	Impfungen sind größtenteils sicher und effektiv.
SARS-CoV-2	-	The coronavirus (SARS-CoV-2) is a serious threat to human health.	Das Corona-Virus (SARS-CoV-2) ist eine ernsthafte Bedrohung für die menschliche Gesundheit.

Participants’ attitudes toward the CSI topics (i.e., acceptance or rejection of the scientific consensus) were measured using a five-point scale to rate their agreement with the five claims. The participants were subsequently asked to justify (i.e., data/warrant/backing) their attitude on each claim in an open answer format and to think of possible reasons to change their position (i.e., rebuttals). In the following analysis, we focus on the justifications.

In addition, other potentially influencing variables were assessed: knowledge about NOS [88], religiousness [43,89], and conspiracy ideation [90]. The NOS measure focused on the tentativeness of scientific knowledge (“development” scale) with items like “New findings might change what scientists hold as true” [88]. The original seven items were reduced to six items. The scale measuring religiousness consisted of five items such as “I believe in God” [43,89]. The scale on conspiracy ideation [90] included items like “I think many important things happen in the world, which the public is never informed about” [90]. All of these scales measured agreement on a five-point rating scale.

3.3. Data Analysis

Results of all rating scales were merged to sum scores per scale. The rating items to assess attitudes toward claims concerning the five CSI topics were merged to one sum score for further analyses, representing attitudes toward scientific consensus.

Open answer format responses (i.e., arguments) were analyzed using qualitative content analysis [91] and the software MAXQDA Plus (VERBI Software, 2019, Berlin,

Germany). Components of the analysis are semantic units; every semantic unit was coded once.

As a first step, based on TAP, we deductively derived an operationalization to identify the semantic units within respondents' arguments that can be categorized as justifications (Figure 3). This step was necessary since, even if the open answer format question concretely asked about justifications, some answers contained other argumentative elements or unrelated components.

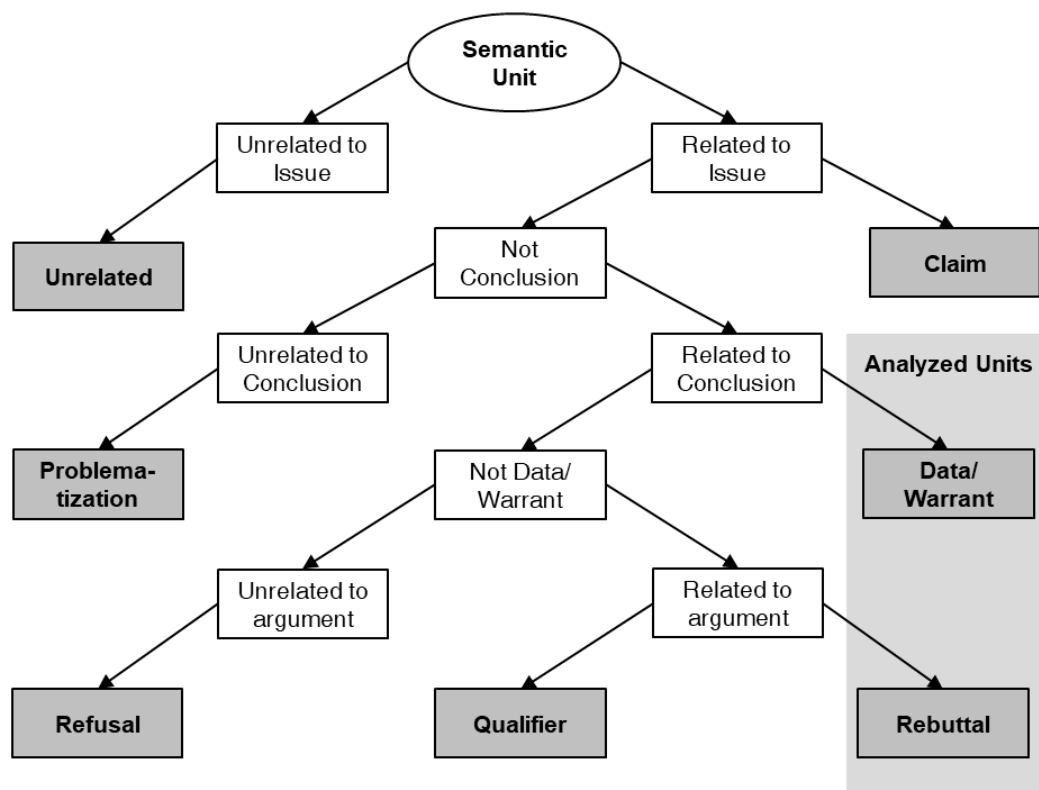


Figure 3. Operationalizing Toulmin's Argumentation Pattern [32] to isolate the justification (i.e., analyzed unit) of the argument proposed in the open answer format as the first step of the analysis.

If the semantic unit named reasons supporting the participant's position concerning the claim (e.g., "The risk of dying from the disease is higher than dying from the vaccine") it was coded as *warrant/data*, since those two argument components, as postulated by Toulmin [32], rarely appear explicitly as two distinct units. In this case, the warrant (i.e., "If the risk of dying from the disease is higher than dying from the vaccine, the vaccine is safe and effective") is left implicit, as is often the case in informal logic [76]. The conceptualization of Toulmin [32] also includes *qualifiers* influencing the magnitude of an argument (e.g., "If the vaccine is developed and tested responsibly") and *rebuttals* contradicting the conclusion (e.g., "Some people die from the side effects of vaccines"). For the following analysis of the justifications, the rebuttals were merged with warrant/data as justifications (i.e., analytic unit; Figure 3), because the statement "Some people die from the side effects of vaccines" either justifies or rebuts a participant's position.

If the semantic unit was completely unrelated to the claim, it was coded as *unrelated*. If it was a restatement of the claim captured in the rating scale (e.g., "I think vaccines help") it was coded as *claim*. If the semantic unit was unrelated to the initial statement, e.g., the safety and effectiveness of vaccines, but still related to the issue (e.g., "No one should be forced to be vaccinated") the unit was coded as *problematization*. Semantic units that referred to the initial statement without using any argumentative component (e.g., "Why would I answer that?") were coded as *refusals*.

The first deductive coding step resulted in a majority of answers justifying the statement, as intended in the open question (Table 2). There was no evidence of structural differences between stated argument components for or against scientific consensus. The stated argument components did not depend on the attitude measured.

Table 2. Frequencies (proportions) of argument components among the five CSI topics following the first deductive coding step. The grey row displays the proportion of semantic units identified as justifications.

Argument Component	Evolution	Climate Change	GMF	Vaccination	SARS-CoV-2	Total
claim	5 (1.3%)	13 (3.6%)	12 (3.1%)	13 (3.2%)	3 (0.8%)	46 (2.4%)
data/warrant/rebutt (i.e., justification)	350 (88.8%)	289 (80.1%)	300 (76.5%)	325 (79.9%)	333 (84.1%)	1597 (81.9%)
qualifier	17 (4.3%)	23 (6.4%)	22 (5.6%)	34 (8.4%)	37 (9.3%)	133 (6.8%)
problematization	2 (0.5%)	23 (6.4%)	41 (10.5%)	25 (6.1%)	10 (2.5%)	101 (5.2%)
refusal	6 (1.5%)	2 (0.6%)	6 (1.5%)	4 (1.0%)	5 (1.3%)	23 (1.2%)
unrelated	14 (3.6%)	11 (3.1%)	11 (2.8%)	6 (1.5%)	8 (2.0%)	50 (2.6%)
Total	394 (100%)	361 (100%)	392 (100%)	407 (100%)	396 (100%)	1950 (100%)

The semantic units identified as justifications underwent a second qualitative content analysis to build up the deductive-inductive category system and answer the research questions. Therefore, we started by gathering similar content in fine-grained subcategories and subsequently generalized the categories more and more [91] based on those presented by Lobato and Zimmerman [36]. In this way, it was possible to categorize the justifications based on content and build types of justifications on CSI topics. To improve the objectivity of our category system, a different researcher conducted a second coding on 30 complete data sets (11.3% of complete data sets) [92]. These double coded data sets are a representational sample to encompass the spectrum of the material. Cohen's kappa indicates a substantial intercoder agreement ($\kappa = 0.68$) [92]. Based on a discursive analysis of the coding results, codings were discursively changed when coding errors were identified. This led to increasement of Cohen's kappa ($\kappa = 0.84$) and a refinement of coding descriptions.

The amount and proportion of coded semantic units per type of justification were calculated and compared across the five topics. To analyze relations between types of justifications and other variables, correlations were calculated.

4. Results

The claims reflecting the scientific consensus on the five CSI topics were generally accepted, representing a positive attitude toward these topics. The most accepted claim was evolution (95.3 % agreement), followed by climate change (87.6%), vaccinations (86.0%), SARS-CoV-2 (82.6%), and finally GMF (57.5%), the most contested claim (Table 3).

Figure 4 displays the deductively-inductively built fine-grained category system to distinguish different types of justifications on CSI. The categorization resulted in five types of justifications, with 25 subcategories. A justification that cannot be falsified or is dependent on individual beliefs belongs to the *subjective* type, and every other justification is *intersubjective*. Subjective justifications refer to normative statements that are grounded in values and beliefs (e.g., *ideology*: "God created all living beings"; *naturalistic fallacy*: "This is not safe, because it is not natural"; *argumentum ad hominem*: "Virologists are not trustworthy"). Intersubjective justifications were further distinguished into those referring to specific data to support the claim (*evidential*) or referring to a third entity as an authority (*deferential*). The mere mention of "evidence" did not count as an evidential justification but

was categorized as a reference to a body of knowledge and therefore as deferential. The determining differentiation between evidential and deferential justifications was their specificity; while deferential justifications refer to a rather general body or lack of knowledge about the topic, evidential justifications are quite focused on the single CSI.

Table 3. Frequency (proportion) of acceptance of scientific consensus concerning each CSI topic.

CSI Topic	Rejection	Undecided	Acceptance	Total
Evolution	13 (3.4 %)	5 (1.3 %)	361 (95.3 %)	379 (100 %)
Climate Change	17 (5.3 %)	23 (7.1 %)	282 (87.6 %)	322 (100 %)
GMF	61 (21.3 %)	61 (21.3 %)	165 (57.5 %)	287 (100 %)
Vaccination	26 (9.4 %)	13 (4.7 %)	239 (86.0 %)	278 (100 %)
SARS-CoV-2	29 (10.7 %)	18 (6.7 %)	223 (82.6 %)	270 (100 %)

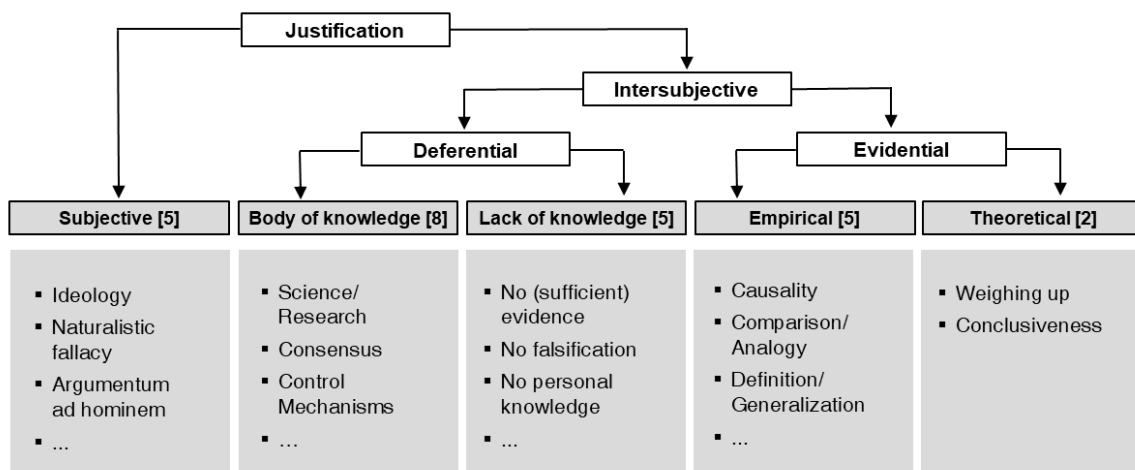


Figure 4. Category system of justification types concerning attitudes toward claims on CSI topics. Number of subcategories per category is in square brackets.

Deferential justifications were further divided into justifications referring to a *body of knowledge* (e.g., *science/research*: “That was proven by science”; *consensus*: “Almost all scientist agree on it”; *control mechanisms*: “There is a strict and transparent approval procedure for vaccinations”) or a *lack of knowledge* (e.g., *no (sufficient) evidence*: “We don’t know enough about it”; *no falsification*: “To date, there is no evidence against it”; *no personal knowledge*: “I don’t know enough about this”). The evidential justifications were categorized as either *empirical* or *theoretical* justifications. While empirical justifications referred to verifiable real-world phenomena (e.g., *causality*: “As shown by the eradication of smallpox”; *comparison/analogy*: “SARS-CoV-2 is not more dangerous than the flu”; *definition/generalization*: “This is the case, since we have a global pandemic”), the theoretical justifications drew conclusions, weighed up, or referred to conclusiveness (e.g., *weighing up*: “The risk of dying from the sickness is higher than dying from the vaccine”; *conclusiveness*: “This is a conclusive explanation”).

All identified justification types were identified across all five topics. However, some justification types were more common depending on the particular CSI topic addressed. While subjective, empirical, and theoretical justifications tended to be rather topic-specific, deferential justifications appeared with a similar frequency across most of the topics (Table 4). Therefore, references to a body or lack of knowledge were used quite similarly across the different CSI. However, on the safety of GMF, the most contested statement, a

comparably high number of justifications refer to a lack of knowledge. In contrast, the claims with the highest acceptance rates, i.e., anthropogenic climate change, evolution, and vaccination, were more frequently connected with justifications referring to third entities or vaguely defined bodies of knowledge such as “studies” or “evidence”.

Table 4. Proportions of justification types across the CSI topics.

Type of Justification	Evolution	Climate Change	GMF	Vaccination	SARS-CoV-2	N_{Total}
Subjective	0.4%	1.4%	7.3%	4.3%	3.6%	66
Deferential: Body of knowledge	49.1%	55.2%	25.3%	45.9%	26.1%	644
Deferential: Lack of knowledge	9.1%	5.9%	38.0%	6.5%	13.8%	230
Evidential: Theoretical	17.1%	6.2%	3.0%	12.3%	1.5%	132
Evidential: Empirical	20.6%	31.4%	26.3%	31.1%	55.0%	526
N_{Total}	350	290	300	325	333	1598

Compared with theoretical justifications, empirical justifications were far more common. However, this varied across the topics; while justifications concerning evolution relied almost equally on theoretical considerations and real-world observations, positions on SARS-CoV-2 were more frequently justified by empirical justifications.

Subjective justifications were the least common justification type, with the topic of GMF showing the highest proportion of subjective justifications, while almost no respondents gave subjective justifications in the contexts of evolution and anthropogenic climate change. The most frequent justification type overall was reference to a body of knowledge. This type was especially common when justifying attitudes on anthropogenic climate change, evolution, and vaccination.

In most cases, the acceptance of claims concerning the five different CSI topics did not correlate significantly with the identified type of justification (Table 5). However, the use of subjective justifications is negatively correlated to the acceptance of four of the CSI topics with a weak effect. The claim about GMF is the only one without a significant correlation to one of the justification types. Additionally, the acceptance of the effectiveness and safety of vaccines is significantly and weakly related to the use of deferential justifications referring to a body of knowledge.

Table 5. Correlation after Pearson between justification type and acceptance of scientific consensus on each topic. $N = 398$, * $p < 0.05$, ** $p < 0.01$.

Justification Type	Evolution	Climate Change	GMF	Vaccination	SARS-CoV-2
Subjective	−0.141 **	−0.175 **	−0.064	−0.186 **	−0.153 *
Deferential: Body of Knowledge	0.044	0.060	0.114	0.128 *	0.110
Deferential: Lack of Knowledge	0.045	0.015	−0.048	−0.008	0.044
Evidential: Theoretical	0.059	−0.003	0.057	−0.015	−0.026
Evidential: Empirical	0.056	−0.036	−0.074	−0.016	0.053

In general, the participants were not very religious ($M = 1.62$; $SD = 1.05$, score range: 1–5), were partly drawn to conspiracy theories ($M = 2.41$; $SD = 0.88$, score range: 1–5), and showed a high knowledge about NOS ($M = 4.69$; $SD = 0.44$, score range: 1–5).

A significant positive and strong correlation between the general acceptance of scientific consensus and knowledge about NOS ($N = 252$; $r = 0.558$; $p < 0.01$) was identified. Religiousness ($N = 254$; $r = -0.469$; $p < 0.01$) and conspiracy ideation ($N = 258$; $r = -0.655$; $p < 0.01$) correlated significantly negatively with the acceptance of scientific consensus with a medium (religiousness) to strong (conspiracy ideation) effect size.

Correlations of these variables with different types of justification were not significant in most cases (Table 6). Solely the use of subjective justifications (e.g., natural fallacy) correlated positively and weakly with the rejection of scientific consensus as well as negatively and weakly with conspiracy ideation. References to a body of knowledge correlated with the acceptance of scientific consensus with a weak effect. Furthermore, religiousness correlated weakly with the use of empirical justifications.

Table 6. Correlations between justification type and knowledge about NOS, religiousness, conspiracy ideation, and general acceptance of scientific consensus operationalized by the mean of acceptance of the claims on the five CSI topics. $N = 398$, * $p < 0.05$, ** $p < 0.01$.

Justification Type	NOS	Religiousness	Conspiracy Ideation	Acceptance of Scientific Consensus
Subjective	−0.045	0.029	0.180 **	−0.185 **
Deferential: Body of Knowledge	0.035	−0.121	0.102	0.102 *
Deferential: Lack of Knowledge	−0.024	0.009	0.030	0.020
Evidential: Theoretical	0.011	0.047	−0.036	0.007
Evidential: Empirical	0.032	0.184 **	0.079	−0.009

5. Discussion

The relatively high agreement with the claims on the different CSI indicates that most citizens who responded to the survey accept the respective scientific consensus. However, while evolution as the explanation for the variety of life forms is accepted by more than 95% of the sample, only 57.5% agreed with the safety of GMF, the claim with the highest frequency of rejection and uncertainty. About 10% disagreed with the claims about the effectiveness and safety of vaccines and the health threat of SARS-CoV-2. Analysis of justifications resulted in five types of justifications for claims on CSI, each with several subtypes (RQ1). Justification types seem to be partly topic specific (RQ2) and in most cases are unrelated to whether the claim on a CSI was accepted or rejected (RQ3), as well as to variables like NOS, religiousness, and conspiracy ideation (RQ4).

5.1. Justification Types in the Field of Controversial Science Issues (RQ1)

To identify types of justifications in the field of CSI, we applied a deductive-inductive approach based on an existing justification coding scheme [36]. We identified subjective justifications that have been described before [36], sometimes referred to as normative justifications [27]. This type relies on individual spiritual, political, or ideological beliefs as well as on reasoning fallacies like *argumentum ad hominem*.

All justifications that could be identified as intersubjective formed a group that was further categorized. The distinction between references to a third entity (i.e., deferential) and references to the subject of discussion itself (i.e., evidential) was drawn from previous research [36] and applied to the data in this study. This common distinction can also be found in Shtulman [83].

However, taking a closer look at the deferential justifications, we distinguished references to a body of knowledge (e.g., “There is evidence for x”) from references to a lack of knowledge (e.g., “There is no evidence”). Another step toward more fine-grained categories was the distinction within the evidential category between empirical and theoretical

justifications. Empirical justifications rely on real-world phenomena or precisely named and therefore provable data (e.g., correlation: “There is a positive correlation between greenhouse gas emissions and rising global temperature”), while theoretical justifications include a warrant to support the conclusion (e.g., cost risk calculation: “Even if climate change is not anthropogenic, we should assume it is. Better safe than sorry”). This categorization of the evidential justifications as either empirical or theoretical is therefore aligned with the distinction between data and warrant in TAP [32]. Furthermore, both types of evidential justifications share commonalities with components of scientific reasoning, e.g., the subskill of interpreting data [20] or abductive reasoning [93]. It would be worth investigating to what extent these types of evidential justifications align with the epistemic dimension of scientific reasoning as described by Osborne [21], referring to the questions “How do we know or how can we be certain?” [21] (p. 270).

Clearly, evidential justifications that refer to the CSI topic under consideration itself are highly topic-dependent. Due to the high diversity of SSI [70], a further generalization of this type of justification is challenging. One step that enabled the categorization into justification types was the focus on CSI as a special variant of SSI. Following Kolstø [94], who defined the field of risk-based SSI, and Borgerding and Dagistan [11] (see Figure 1), who differentiated between different fields as foundations for SSI, the theoretical clarification of the field of CSI as well as the resulting category system may help to further clarify the different fields within the broad topic of SSI. This is likely necessary for a finer analysis of justifications that could perhaps be field-specific.

5.2. Topic-Specific Justifications (RQ2)

Despite the field-specific scope of the category system, indicated by the occurrence of all five justification types in all five CSI topics, the results show frequency differences among justifications concerning the five CSI. This finding supports earlier results with a similar methodological design [36], while results of studies investigating SSR instead suggest consistency of the SSR framework across different SSI contexts [8,69]. However, this contrast may be resolved by seeing the SSR framework as a field-specific tool that is applicable to different topics of SSI, comparable with the category system for the field of CSI presented here. Toulmin [32] has already emphasized the field-specificity of arguments.

Whereas subjective and evidential justifications appear to be more topic-specific, the most general justification types seem to be deferential justifications referring to a body or lack of knowledge, either personal or related to the scientific field. In fact, the vast majority of deferential justifications refer to the scientific field. However, as the participants were aware that they were part of a scientific survey, they may have tried to use appropriate and convincing arguments. Laypeople are often capable of using “public scientific arguments” [25].

5.3. Relationship between Acceptance of CSI and the Use of Different Justifications (RQ3)

Generally, correlations between the use of certain justifications and the acceptance of the scientific consensus on the different CSI were weak. Still, the use of subjective justifications correlated with a rejection of the scientific consensus on most CSI, except for the safety of genetically modified food. One possible explanation is that all kinds of fallacies (i.e., argumentum ad populum, argumentum ad hominem, naturalistic fallacy) are subjective justifications. This fallacious argumentation is known to be rather common when arguing against a scientific consensus [95]. Despite the only small number of subjective justifications in total, these correlations suggest that subjective justifications are more frequently formulated if people reject the scientific consensus on a CSI.

Deferential and evidential justifications seem to appear for both acceptance and rejection of the scientific consensus, indicated by the insignificant correlations between the use of these justifications and acceptance of the scientific consensus on the five CSI. The only exception is a significant and weak correlation between reference to a body of

knowledge and acceptance of the effectiveness and safety of vaccinations, indicating less frequent use of this argument when being skeptical about vaccinations.

5.4. Relationship between NOS, Religiousness, and Conspiracy Ideation with the Use of Different Justifications (RQ4)

Concerning the relationship between justification types and other variables, increased knowledge about NOS did not correlate with a certain type of justification, an observation made previously concerning NOS and the structural quality of arguments [47]. Nevertheless, NOS is known to be able to positively influence argumentation skills on SSI topics [85,96].

While Lobato and Zimmerman [36] noted that justification strategies appear highly heterogeneous within an individual's argumentation, our research demonstrated that even across the spectrum of science rejection and acceptance, all different kinds of justifications appear. This is consistent with previous findings that point out similarities in argumentation on supernatural beliefs and scientific knowledge [83]. However, significant correlations indicate that reference to a body of knowledge is more likely when accepting the scientific consensus, while subjective justifications are more frequent in argumentations against the scientific consensus.

Furthermore, subjective justifications are more frequent in people with high conspiracy ideation. Religiousness correlated weakly and positively with the use of empirical justifications, suggesting that religiousness is not necessarily an obstacle to reasoning on scientific topics [43].

6. Conclusions and Outlook

The task of fostering reasoning and argumentation competency goes beyond formal education in school and university [4]. In general, citizens are expected to employ evidence-based reasoning on issues grounded in science to make decisions in their personal lives and in public policy [97]. People often have difficulty evaluating evidence, which is problematic for informal reasoning on public policy and personal choices [4]. One crucial reason that these everyday reasoning tasks are difficult is the easy generation of causal explanation and their resistance [4,98].

To equip citizens with the ability to weigh up arguments and evaluate evidence, a first step is knowledge about the different types of justifications they provide for their attitudes concerning certain CSI. The category system reflecting justification types provides insight into the diversity of argumentation patterns and can inform teachers and pre-service teachers about potential attitudes and justifications on CSI that they might encounter in their lessons. Previous studies emphasized the importance of the inclusion of multidisciplinary perspectives when negotiating complex societal issues like CSI [7,35]. This approach can be informed by the category system, which was built upon a wide variety of different justifications from a heterogeneous online sample. It could furthermore be a helpful tool for fostering *science media literacy*, described by Höttecke and Allchin [99] as a crucial goal of science education in the age of social media [99].

Moreover, the presented category system lays the groundwork for further research in this area. On one hand, it will be the starting point for similar research in formal education. On the other hand, knowledge about justification types and how they differ across different contexts enables the ability to choose the best contexts to integrate into science education contexts.

Additionally, the results may inform science communication researchers and practitioners about the acceptance of the scientific consensus on different CSI topics and common justifications in these contexts. This is important, since even media reports often have problems handling scientific information [19].

In future research, the fine-grained assessment of general attitudes toward SSI brought forward by Klaver and Walma van der Molen [67] could be combined with the method of measuring justifications toward scientific consensus on specific CSI proposed in this article to shed more light on the different justification types. Furthermore, a research

design integrating a task on SSR would be beneficial, e.g., by using the QuASSR [65]. In general, further investigation of the category system and its justification types should include steps of further validation [100] as well as argumentation in a broader discussion context, as has been suggested by several scholars [32,33,72]. The current study involved a random sample recruited within social networks to collect a wide variety of justifications for creating the category system. However, this sampling led to a high dropout rate and lacks representativeness of the quantified results. Future studies may apply the category system to samples within controlled environments. Another important next step is the theoretical and empirical investigation of the alignment of scientific reasoning and informal reasoning on CSI and SSI.

The novel term CSI could—following further theoretical and empirical clarification—help bridge the gap between the mostly separated research areas of science education and science communication [18].

Author Contributions: Conceptualization, A.B.; methodology, A.B.; validation, A.B., L.M. and A.U.z.B.; formal analysis, L.M.; investigation, L.M.; resources, A.B. and A.U.z.B.; data curation, L.M. and A.B.; writing—original draft preparation, L.M. and A.B.; writing—review and editing, A.B., A.U.z.B. and L.M.; visualization, A.B.; supervision, A.B. and A.U.z.B.; project administration, A.B. and A.U.z.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: The respondents agreed to data use for research.

Data Availability Statement: The datasets are not publicly available.

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

References

1. OECD. OECD Future of Education and Skills 2030. OECD Learning Compass 2030. A Series of Concept Notes. Available online: http://www.oecd.org/education/2030-project/contact/OECD_Learning_Compass_2030_Concept_Note_Series.pdf (accessed on 18 July 2021).
2. Gauchat, G. Politicization of Science in the Public Sphere: A Study of Public Trust in the United States, 1974 to 2010. *Am. Sociol. Rev.* **2012**, *77*, 167–187. [CrossRef]
3. Liu, D.W.C. Science Denial and the Science Classroom. *CBE Life Sci. Educ.* **2012**, *11*, 129–134. [CrossRef] [PubMed]
4. Shah, P.; Michal, A.; Ibrahim, A.; Rhodes, R.; Rodriguez, F. What makes everyday scientific reasoning so challenging? *Psychol. Learn. Motiv.* **2017**, *66*, 251–299. [CrossRef]
5. Sadler, T.D. Informal reasoning regarding socioscientific issues: A critical review of research. *J. Res. Sci. Teach.* **2004**, *41*, 513–536. [CrossRef]
6. Zeidler, D.L.; Sadler, T.D.; Simmons, M.L.; Howes, E.V. Beyond STS: A research-based framework for socioscientific issues education. *Sci. Educ.* **2005**, *89*, 357–377. [CrossRef]
7. Romine, W.L.; Sadler, T.D.; Dauer, J.M.; Kinslow, A. Measurement of socio-scientific reasoning (SSR) and exploration of SSR as a progression of competencies. *Int. J. Sci. Educ.* **2020**, *42*, 2981–3002. [CrossRef]
8. Sadler, T.D.; Barab, S.A.; Scott, B. What do students gain by engaging in socioscientific inquiry? *Res. Sci. Educ.* **2007**, *37*, 371–391. [CrossRef]
9. Zeidler, D.; Sadler, T.; Applebaum, S.; Callahan, B. Advancing reflective judgment through socioscientific issues. *J. Res. Sci. Teach.* **2009**, *46*, 74–101. [CrossRef]
10. Rutjens, B.T.; van der Linden, S.; van der Lee, R. Science skepticism in times of COVID-19. *Group Process. Intergr. Relat.* **2021**, *24*, 276–283. [CrossRef]
11. Borgerding, L.A.; Dagistan, M. Preservice science teachers' concerns and approaches for teaching socioscientific and controversial issues. *J. Sci. Teach. Educ.* **2018**, *29*, 283–306. [CrossRef]
12. McComas, W.F. Controversial Science Issues. In *The Language of Science Education*, 1st ed.; McComas, W.F., Ed.; SensePublishers: Rotterdam, The Netherlands, 2014; p. 26. [CrossRef]
13. Hornsey, M.J.; Fielding, K.S. Attitude roots and Jiu Jitsu persuasion: Understanding and overcoming the motivated rejection of science. *Am. Psychol.* **2017**, *72*, 459–473. [CrossRef] [PubMed]
14. Sadler, T.D.; Zeidler, D.L. Scientific literacy, PISA, and socioscientific discourse: Assessment for progressive aims of science education. *J. Res. Sci. Teach.* **2009**, *46*, 909–921. [CrossRef]
15. Burns, T.W.; O'Connor, D.J.; Stocklmayer, S.M. Science communication: A contemporary definition. *Public Underst. Sci.* **2003**, *12*, 183–202. [CrossRef]

16. Eastwood, J.L.; Schlegel, W.M.; Cook, K.L. Effects of an Interdisciplinary Program on Students' Reasoning with Socioscientific Issues and Perceptions of Their Learning Experiences. In *Socio-Scientific Issues in the Classroom—Teaching, Learning and Research*, 1st ed.; Sadler, T.D., Ed.; Springer: Dordrecht, The Netherlands, 2011; pp. 89–126. [CrossRef]
17. Upmeier zu Belzen, A.; Beniermann, A. Naturwissenschaftliche Grundbildung im Fächerkanon der Schule. *Z. Padagog.* **2020**, *66*, 642–665.
18. Baram-Tsabari, A.; Osborne, J. Bridging science education and science communication research. *J. Res. Sci. Teach.* **2015**, *52*, 135–144. [CrossRef]
19. Bromme, R.; Goldman, S.R. The public's bounded understanding of science. *Educ. Psychol.* **2014**, *49*, 59–69. [CrossRef]
20. Hartmann, S.; Upmeier zu Belzen, A.; Krüger, D.; Pant, H.A. Scientific reasoning in Higher Education: Constructing and Evaluating the Criterion-Related Validity of an Assessment of Preservice Science Teachers' Competencies. *Z. Psychol.* **2015**, *223*, 47–53. [CrossRef]
21. Osborne, J. The 21st century challenge for science education: Assessing scientific reasoning. *Think. Ski. Creat.* **2013**, *10*, 265–279. [CrossRef]
22. Means, M.L.; Voss, J.F. Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cogn. Instr.* **1996**, *14*, 139–178. [CrossRef]
23. Jiménez-Aleixandre, M.P.; Erduran, S. Argumentation in Science Education: An Overview. In *Argumentation in Science Education: Perspectives from Classroom-Based Research*, 1st ed.; Erduran, S., Jiménez-Aleixandre, M.P., Eds.; Springer: Dordrecht, The Netherlands, 2007; Volume 35, pp. 3–27. [CrossRef]
24. Osborne, J.; Erduran, S.; Simon, S. Enhancing the quality of argumentation in school science. *J. Res. Sci. Teach.* **2004**, *41*, 994–1020. [CrossRef]
25. Endres, D. Science and public participation: An analysis of public scientific argument in the Yucca Mountain controversy. *Environ. Commun.* **2009**, *3*, 49–75. [CrossRef]
26. Gresch, H.; Schwanewedel, J. Argumentieren als naturwissenschaftliche Praktik. In *Biologiedidaktische Forschung: Erträge für die Praxis*, 1st ed.; Groß, J., Hammann, M., Schmiemann, P., Zabel, J., Eds.; Springer Spektrum: Berlin/Heidelberg, Germany, 2019; pp. 167–185. [CrossRef]
27. Jafari, M.; Meisert, A. Activating students' argumentative resources on socioscientific issues by indirectly instructed reasoning and negotiation processes. *Res. Sci. Educ.* **2019**, 1–22. [CrossRef]
28. Zohar, A.; Nemet, F. Fostering Students' Knowledge and Argumentation Skills Through Dilemmas in Human Genetics. *J. Res. Sci. Teach.* **2002**, *39*, 35–62. [CrossRef]
29. Khishfe, R. Explicit nature of science and argumentation instruction in the context of socioscientific issues: An effect on student learning and transfer. *Int. J. Sci. Educ.* **2014**, *36*, 974–1016. [CrossRef]
30. Sjöström, J.; Eilks, I. Reconsidering different visions of scientific literacy and science education based on the concept of Bildung. In *Cognition, Metacognition, and Culture in STEM Education*, 1st ed.; Dori, Y.J., Mevarech, Z.R., Baker, D.R., Eds.; Springer: Cham, Switzerland, 2018; pp. 65–88. [CrossRef]
31. Sampson, V.; Clark, D.B. Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Sci. Educ.* **2008**, *92*, 447–472. [CrossRef]
32. Toulmin, S.E. *The Uses of Argument*; Cambridge University Press: Cambridge, UK, 2003; (Original Publication in 1958).
33. Lazarou, D.; Erduran, S. "Evaluate What I Was Taught, Not What You Expected Me to Know": Evaluating Students' Arguments Based on Science Teachers' Adaptations to Toulmin's Argument Pattern. *J. Sci. Teach. Educ.* **2021**, *32*, 306–324. [CrossRef]
34. Henderson, B.J.; McNeill, K.L.; Gonzalez-Howard, M.; Close, K.; Evans, M. Key challenges and future directions for educational research on scientific argumentation. *J. Res. Sci. Teach.* **2017**, *55*, 5–18. [CrossRef]
35. Garrecht, C.; Reiss, M.J.; Harms, U. 'I wouldn't want to be the animal in use nor the patient in need'—The role of issue familiarity in students' socioscientific argumentation. *Int. J. Sci. Educ.* **2021**, 1–22. [CrossRef]
36. Lobato, E.J.; Zimmerman, C. Examining how people reason about controversial scientific topics. *Think. Reason.* **2019**, *25*, 231–255. [CrossRef]
37. Sadler, T.D.; Zeidler, D.L. The significance of content knowledge for informal reasoning regarding socioscientific issues: Applying genetics knowledge to genetic engineering issues. *Sci. Educ.* **2005**, *89*, 71–93. [CrossRef]
38. Drummond, C.; Fischhoff, B. Individuals with greater science literacy and education have more polarized beliefs on controversial science topics. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 9587–9592. [CrossRef]
39. Jang, S.M. Seeking congruency or incongruency online? Examining selective exposure to four controversial science issues. *Sci. Commun.* **2014**, *36*, 143–167. [CrossRef]
40. Evagorou, M.; Dillon, J. Introduction: Socio-scientific Issues as Promoting Responsible Citizenship and the Relevance of Science. In *Science Teacher Education for Responsible Citizenship*, 1st ed.; Evagorou, M., Nielsen, J., Dillon, J., Eds.; Contemporary Trends and Issues in Science Education; Springer: Cham, Switzerland, 2020; pp. 1–11. [CrossRef]
41. Zeidler, D.L. Socioscientific issues as a curriculum emphasis: Theory, research and practice. In *Handbook of Research on Science Education*; Lederman, N.G., Abell, S.K., Eds.; Routledge: New York, NY, USA, 2014; Volume 2, pp. 697–726.
42. Zeidler, D.L.; Herman, B.C.; Sadler, T.D. New directions in socioscientific issues research. *Discipl. Interdiscip. Sci. Educ. Res.* **2019**, *1*, 1–9. [CrossRef]

43. Beniermann, A. *Evolution—von Akzeptanz und Zweifeln—Empirische Studien über Einstellungen zu Evolution und Bewusstsein*, 1st ed.; Springer Fachmedien: Wiesbaden Germany, 2019; pp. 1–469.
44. Kahan, D.M. Climate-science communication and the measurement problem. *Political Psychol.* **2015**, *36*, 1–43. [CrossRef]
45. Eagly, A.H.; Chaiken, S. *The Psychology of Attitudes*; Harcourt Brace Jovanovich College Publishers: Fort Worth, TX, USA, 1993.
46. Betsch, C.; Schmid, P.; Heinemeier, D.; Korn, L.; Holtmann, C.; Böhm, R. Beyond confidence: Development of a measure assessing the 5C psychological antecedents of vaccination. *PLoS ONE* **2018**, *13*, e0208601. [CrossRef]
47. Christenson, N.; Rundgren, S.N.C. A framework for teachers' assessment of socio-scientific argumentation: An example using the GMO issue. *J. Biol. Educ.* **2015**, *49*, 204–212. [CrossRef]
48. Graf, D.; Soran, H. Einstellung und Wissen von Lehramtsstudierenden zur Evolution—Ein Vergleich zwischen Deutschland und der Türkei. In *Evolutionstheorie—Akzeptanz und Vermittlung im Europäischen Vergleich*, 1st ed.; Graf, D., Ed.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 141–161. [CrossRef]
49. Dunlap, R.E.; McCright, A.M. A widening gap: Republican and Democratic views on climate change. *Environ. Sci. Policy* **2008**, *50*, 26–35. [CrossRef]
50. Kahan, D.M.; Wittlin, M.; Peters, E.; Slovic, P.; Ouellette, L.L.; Braman, D.; Mandel, G.N. The tragedy of the risk-perception commons: Culture conflict, rationality conflict, and climate change. *Temple Univ. Leg. Stud. Res. Pap.* **2011**, *26*, 1–31. [CrossRef]
51. Scott, S.E.; Inbar, Y.; Wirz, C.D.; Brossard, D.; Rozin, P. An overview of attitudes toward genetically engineered food. *Annu. Rev. Nutr.* **2018**, *38*, 459–479. [CrossRef] [PubMed]
52. Blancke, S.; Van Breusegem, F.; De Jaeger, G.; Braeckman, J.; Van Montagu, M. Fatal attraction: The intuitive appeal of GMO opposition. *Trends Plant Sci.* **2015**, *20*, 414–418. [CrossRef] [PubMed]
53. Rutjens, B.T.; Sutton, R.M.; van der Lee, R. Not all skepticism is equal: Exploring the ideological antecedents of science acceptance and rejection. *Pers. Soc. Psychol.* **2018**, *44*, 384–405. [CrossRef]
54. Rutjens, B.T.; van der Lee, R. Spiritual skepticism? Heterogeneous science skepticism in the Netherlands. *Public Underst. Sci.* **2020**, *29*, 335–352. [CrossRef]
55. Marques, M.D.; Kerr, J.R.; Williams, M.N.; Ling, M.; McLennan, J. Associations between conspiracism and the rejection of scientific innovations. *Public Underst. Sci.* **2021**, 1–14. [CrossRef]
56. Weisberg, D.S.; Landrum, A.R.; Hamilton, J.; Weisberg, M. Knowledge about the nature of science increases public acceptance of science regardless of identity factors. *Public Underst. Sci.* **2021**, *30*, 120–138. [CrossRef]
57. Shaw, V.F. The cognitive processes in informal reasoning. *Think. Reason.* **1996**, *2*, 51–80. [CrossRef]
58. Evans, J.S.B.T.; Thompson, V.A. Informal reasoning: Theory and method. *Can. J. Exp. Psychol.* **2004**, *58*, 69–74. [CrossRef] [PubMed]
59. Evans, J.S.B.T. Logic and human reasoning: An assessment of the deduction paradigm. *Psychol. Bull.* **2002**, *128*, 978–996. [CrossRef] [PubMed]
60. Kuhn, D. Connecting scientific and informal reasoning. *Merrill-Palmer Q.* **1993**, *39*, 74–103.
61. Wu, Y.T.; Tsai, C.C. High school students' informal reasoning on a socio-scientific issue: Qualitative and quantitative analyses. *Int. J. Sci. Educ.* **2007**, *29*, 1163–1187. [CrossRef]
62. Kolstø, S.D. Scientific literacy for citizenship: Tools for dealing with the science dimension of controversial socioscientific issues. *Sci. Educ.* **2001**, *85*, 291–310. [CrossRef]
63. Zeidler, D.L.; Lewis, J. Unifying themes in moral reasoning on socioscientific issues and discourse. In *The Role of Moral Reasoning on Socioscientific Issues and Discourse in Science Education*, 1st ed.; Zeidler, D.L., Ed.; Springer: Dordrecht, The Netherlands, 2003; Volume 19, pp. 289–306.
64. Zeidler, D.L.; Nichols, B.H. Socioscientific issues: Theory and practice. *J. Elem. Sci. Educ.* **2009**, *21*, 49–58. [CrossRef]
65. Romine, W.L.; Sadler, T.D.; Kinslow, A.T. Assessment of scientific literacy: Development and validation of the Quantitative Assessment of Socio-Scientific Reasoning (QuASSR). *J. Res. Sci. Teach.* **2017**, *54*, 274–295. [CrossRef]
66. Eggert, S.; Bögeholz, S. Students' use of decision-making strategies with regard to socioscientific issues: An application of the Rasch partial credit model. *Sci. Educ.* **2010**, *94*, 230–258. [CrossRef]
67. Klaver, L.T.; Walma van der Molen, J.H. Measuring Pupils' Attitudes Towards Socioscientific Issues. *Sci. Educ.* **2021**, *30*, 317–344. [CrossRef]
68. Yang, F.Y.; Anderson, O.R. Senior high school students' preference and reasoning modes about nuclear energy use. *Int. J. Sci. Educ.* **2003**, *25*, 221–244. [CrossRef]
69. Sadler, T.D.; Klosterman, M.L.; Topcu, M.S. Learning Science Content and Socio-scientific Reasoning Through Classroom Explorations of Global Climate Change. In *Socio-Scientific Issues in the Classroom—Teaching, Learning and Research*, 1st ed.; Sadler, T.D., Ed.; Springer: Dordrecht, The Netherlands, 2011; pp. 45–77. [CrossRef]
70. Cian, H. The influence of context: Comparing high school students' socioscientific reasoning by socioscientific topic. *Int. J. Sci. Educ.* **2020**, *42*, 1503–1521. [CrossRef]
71. Topcu, M.S.; Sadler, T.D.; Yilmaz-Tuzun, O. Preservice science teachers' informal reasoning about socioscientific issues: The influence of issue context. *Int. J. Sci. Educ.* **2010**, *32*, 2475–2495. [CrossRef]
72. Driver, R.; Newton, P.; Osborne, J. Establishing the norms of scientific argumentation in classrooms. *Sci. Educ.* **2000**, *84*, 287–312. [CrossRef]
73. Siegel, H. Why should educators care about argumentation? *Inform. Log.* **1995**, *17*, 159–176. [CrossRef]

74. Schwarz, B.; Glassner, A. The blind and the paralytic: Supporting argumentation in everyday and scientific issues. In *Arguing to Learn: Confronting Cognitions in Computer-Supported Collaborative Learning Environments*, 1st ed.; Andriessen, J., Baker, M., Suthers, D., Eds.; Springer: Dordrecht, The Netherlands, 2003; Volume 1, pp. 227–260. [CrossRef]
75. Sandoval, W.A.; Millwood, K. The quality of students' use of evidence in written scientific explanations. *Cogn. Instr.* **2005**, *23*, 23–55. [CrossRef]
76. Paglieri, F. *Coding between the Lines: On the Implicit Structure of Arguments and Its Import for Science Education*; ISTC-CNR Roma; University of Siena: Siena, Italy, 2006.
77. Erduran, S.; Simon, S.; Osborne, J. TAPping into argumentation: Developments in the application of Toulmin's Argument Pattern for studying science discourse. *Sci. Educ.* **2004**, *88*, 915–933. [CrossRef]
78. Koomen, H.M.; Rodriguez, E.; Hoffman, A.; Petersen, C.; Oberhauser, K. Authentic science with citizen science and student-driven science fair projects. *Sci. Educ.* **2018**, *102*, 593–644. [CrossRef]
79. McNeill, K.L.; Lizotte, D.J.; Krajcik, J.; Marx, R.W. Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *J. Learn. Sci.* **2006**, *15*, 153–191. [CrossRef]
80. Kelly, G.J.; Druker, S.; Chen, C. Students' reasoning about electricity: Combining performance assessments with argumentation analysis. *Int. J. Sci. Educ.* **1998**, *20*, 849–871. [CrossRef]
81. Kelly, G.; Takao, A. Epistemic levels in argument: An analysis of university oceanography students' use of evidence in writing. *Sci. Educ.* **2002**, *86*, 314–342. [CrossRef]
82. Dawson, V.; Carson, K. Introducing argumentation about climate change socioscientific issues in a disadvantaged school. *Res. Sci. Educ.* **2018**, *50*, 863–883. [CrossRef]
83. Shtulman, A. Epistemic similarities between students' scientific and supernatural beliefs. *J. Educ. Psychol.* **2013**, *105*, 199–212. [CrossRef]
84. Khishfe, R. Relationship between nature of science understandings and argumentation skills: A role for counterargument and contextual factors. *J. Res. Sci. Teach.* **2012**, *49*, 489–514. [CrossRef]
85. Simonneaux, L. Argumentation in socio-scientific contexts. In *Argumentation in Science Education: Perspectives from Classroom-Based Research*, 1st ed.; Erduran, S., Jiménez-Aleixandre, M.P., Eds.; Springer: Dordrecht, The Netherlands, 2007; Volume 35, pp. 179–199. [CrossRef]
86. Basel, N.; Harms, U.; Prechtel, H.; Weiß, T.; Rothgangel, M. Students' arguments on the science and religion issue: The example of evolutionary theory and Genesis. *J. Biol. Educ.* **2014**, *48*, 179–187. [CrossRef]
87. Agle, J. Assessing changes in US public trust in science amid the COVID-19 pandemic. *Public Health* **2020**, *183*, 122–125. [CrossRef] [PubMed]
88. Urhahne, D.; Kremer, K.; Mayer, J. Conceptions of the nature of science—are they general or context specific? *Int. J. Sci. Math. Educ.* **2011**, *9*, 707–730. [CrossRef]
89. Beniermann, A.; Kuschmierz, P.; Pinxten, R.; Aivelo, T.; Bohlin, G.; Brennecke, J.S.; Cebesoy, U.B.; Cvetković, D.; Đorđević, M.; Dvořáková, R.M.; et al. Evolution Education Questionnaire on Acceptance and Knowledge (EEQ) - Standardised and ready-to-use protocols to measure acceptance of evolution and knowledge about evolution in an international context. *Zenodo* **2021**. [CrossRef]
90. Bruder, M.; Haffke, P.; Neave, N.; Nouripanah, N.; Imhoff, R. Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy Mentality Questionnaire. *Front. Psychol.* **2013**, *4*, 1–15. [CrossRef] [PubMed]
91. Mayring, P. *Qualitative Inhaltsanalyse*, 12th ed.; Beltz: Weinheim, Germany, 2015.
92. O'Connor, C.; Joffe, H. Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *Int. J. Qual. Methods* **2020**, *19*, 1–13. [CrossRef]
93. Upmeyer zu Belzen, A.; Engelschalt, P.; Krüger, D. Modeling as Scientific Reasoning—The Role of Abductive Reasoning for Modeling Competence. *Educ. Sci.* **2021**, *11*, 495. [CrossRef]
94. Kolstø, S.D. Patterns in students' argumentation confronted with a risk-focused socio-scientific issue. *Int. J. Sci. Educ.* **2006**, *28*, 1689–1716. [CrossRef]
95. Zeidler, D.L.; Osborne, J.; Erduran, S.; Simon, S.; Monk, M. The role of argument during discourse about socioscientific issues. In *The Role of Moral Reasoning on Socioscientific Issues and Discourse in Science Education*, 1st ed.; Zeidler, D.L., Ed.; Springer: Dordrecht, The Netherlands, 2003; Volume 19, pp. 97–116.
96. Khishfe, R. Explicit Instruction and Student Learning of Argumentation and Nature of Science. *J. Sci. Teach. Educ.* **2021**, *32*, 325–349. [CrossRef]
97. Kuhn, D.; Lerman, D. Yes but: Developing a critical stance toward evidence. *Int. J. Sci. Educ.* **2021**, *43*, 1036–1053. [CrossRef]
98. Glassner, A.; Weinstock, M.; Neuman, Y. Pupils' evaluation and generation of evidence and explanation in argumentation. *Br. J. Educ. Psychol.* **2005**, *75*, 105–118. [CrossRef]
99. Höttecke, D.; Allchin, D. Reconceptualizing nature-of-science education in the age of social media. *Sci. Educ.* **2020**, *104*, 641–666. [CrossRef]
100. American Educational Research Association (AERA); American Psychological Association (APA); National Council on Measurement in Education (NCME). *Standards for Educational and Psychological Testing*; American Educational Research Association: Washington, DC, USA, 2014.

Article

Elementary Students' Reasoning in Drawn Explanations Based on a Scientific Theory

Valeria M. Cabello ^{1,2,*} , Patricia M. Moreira ¹ and Paulina Griño Morales ^{2,3}¹ Facultad de Educación, Pontificia Universidad Católica de Chile, Santiago 7820436, Chile; pmmoreira@uc.cl² Research Center for Integrated Disaster Risk Management (CIGIDEN), ANID/FONDAP/15110017, Pontificia Universidad Católica de Chile, Santiago 7820436, Chile; paulina.grino@uoh.cl³ Escuela de Educación, Universidad de O'Higgins, Rancagua 2841959, Chile

* Correspondence: vmcabello@uc.cl

Abstract: Constructing explanations of scientific phenomena is a high-leverage practice that promotes student understanding. In the context of this study, we acknowledge that children are used to receiving explanations from teachers. However, they are rarely encouraged to construct explanations about the causes and consequences of phenomena. We modified a strategy to elicit and analyze primary students' reasoning based on scientific theory as a methodological advance in learning and cognition. The participants were fourth-graders of middle socioeconomic status in Chile's geographical zone with high seismic risk. They drew explanations about the causes and consequences of earthquakes during a learning unit of eighteen hours oriented toward explanation-construction based on the Tectonic Plates Theory. A constant comparative method was applied to analyze drawings and characterize students' reasoning used in pictorial representations, following the first coding step of the qualitative Grounded Theory approach. The results show the students expressed progressive levels of reasoning. However, several participants expressed explanations based on the phenomena causes even at an early stage of formal learning. More sophisticated reasoning regarding the scientific theory underpinning earthquakes was found at the end of the learning unit. We discuss approaching elementary students' scientific reasoning in explanations based on theory, connected with context-based science education.

Keywords: explanations; scientific reasoning; drawings; science education; earthquakes

Citation: Cabello, V.M.; Moreira, P.M.; Griño Morales, P. Elementary Students' Reasoning in Drawn Explanations Based on a Scientific Theory. *Educ. Sci.* **2021**, *11*, 581. <https://doi.org/10.3390/educsci11100581>

Academic Editors: Moritz Krell, Andreas Vorholzer and Andreas Nehring

Received: 8 August 2021

Accepted: 21 September 2021

Published: 26 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Instructional practices that are central to learning are called high-leverage practices [1]. Constructing explanations based on evidence derived from inquiry processes [2] or underpinned by scientific theories or principles is relevant for mobilizing students' understanding of natural phenomena in science classrooms [3].

Constructing better explanations continuously provides an organizational and educational framework for designing science teaching and learning experiences [4]. Elementary school students' explanation construction has been researched primarily in developed countries, i.e., [5]. Nonetheless, in developing countries, this field of research is in its early years [6]. Moreover, most of the studies in elementary classrooms are based on students' written explanations [6]. For instance, Forbes et al. [5] found that German classrooms supported students' use of evidence to ground claims. At the same time, the teachers gave more robust opportunities to evaluate evidence-based explanations through comparison in the US. Hence, students learned to look for bias in their reasoning by analyzing other students' explanations. In primary school, exploring and fostering students' explanation construction at the same time is difficult because the students are at the entry points to learn the theories, concepts, or principles. They also start developing writing skills and knowing to use the diverse genres in science education [6]. Thus, misinterpreting students' knowledge because of them having diminished writing skills is likely to happen.

Despite the different emphases of the international studies, there is agreement that explanation-construction is a challenging task for students and teachers which requires explicit support from linguistic and conceptual areas [6,7] or distributed scaffolding to help students gradually [8,9]. Linking phenomena with their underlying causes appears to be among students' difficulties in constructing explanations. This process requires scaffolding and reframing the thinking mechanisms to include theories or concepts already existing in the individual's system of theories [9–12]. Indeed, there is a need for research on evaluative approaches to scaffold students' construction of scientific explanations [11].

Scientific explanations constitute a specialized genre of the discipline in the classroom, different from the report, arguments, or other text genres that children might be more familiar with [13,14]. Constructing explanations also involves the development of causal reasoning [15,16], disciplinary-specificities, i.e., [17], and the transformation of the individuals' intuitive theories [18]. This transformation is influenced by formal knowledge [19], which usually occurs in a social dimension of learning in the classroom activity. Additionally, scientific reasoning skills and other cognitive, metacognitive and motivational—social—skills are related to one another [11,20]. Managing all these dimensions is relevant but also challenging for teachers and researchers when they identify the development of the explaining practice and engage primary students in making sense of phenomena [20–24].

The current study focuses on analyzing pictorial representations of a specific phenomenon, earthquakes, in elementary school students to understand better the process of eliciting their causal reasoning through drawn explanations during a learning sequence. The objective was to characterize students' expressed reasoning through drawn explanations. Using drawings for this purpose advances an evaluative approach to younger learners' thinking, who are just learning to write and talk in science. Additionally, analyzing drawings complements the classic methodological trends of verbal and written modes of making meaning. This knowledge is needed to analyze students' reasoning in phenomena underlined by a scientific theory and identify alternative formats to benefit the growing number of students learning science through a foreign language or those with verbal/oral expression difficulties [24].

1.1. Explanation-Construction as a Meaning-Making Process

Creating or sharing meaning in science education involves multimodal languages, experiences, and interactions in the classroom [21]. The students' construction of explanations as a source of expressing their ideas is crucial, as it provides a window to understanding and sensemaking [22]. A teaching approach responsive to meaning-making processes will anticipate students' ideas about phenomena before instruction and then elicit and respond to these ideas during the lesson [23]. The materials and resources are other crucial elements for meaning-making processes [24].

From a sociocultural perspective, students' explanation construction is a strategy for knowledge integration. It is an iterative and collaborative process in which they connect what is already known—by prior instruction or intuitive theories—with their experiences and conceptual elements to give scientific support for certain phenomena. From this perspective, the explanations constructed are learning artifacts rather than products or learning samples [25]. Explanations in the form of pictorial representations are considered in this study to be vehicles for thought, or reasoning artifacts [26] that trigger the creation of meaning [21] and, consequently, turn into steps in the development of precursor models. These are “cognitive schemata compatible with scientifically appropriate knowledge since they are constructed on the basis of certain elements pertinent to scientific models, which have a limited range of application, and which prepare children's thinking for the construction of scientifically appropriate models” [27] (p. 2259).

In cognitive terms, explanation-construction requires a process of reasoning about phenomena [17] that is rarely easy to access as an external observer since it might require the recreation of the “image of the world” of the other, which contains not only concepts,

but the images created through visual thinking [28]. Indeed, even when teachers know their students' initial ideas, it is hard to build on those ideas while teaching to probe their students' reasoning [29].

Even though interpreting and building new ideas based on students' reasoning in the classroom is challenging for teachers, encouraging the students to construct explanations provides an optimal scenario to engage in understanding natural phenomena, such as those related to socio-scientific issues [30,31]. Moreover, these scenarios help them reconstruct their knowledge and reasoning about phenomena relevant to their lives [32,33]. The reasoning process elicited in the classroom is afforded by an interaction between two information processing systems: the individual's intuitive and deliberative thinking [19]. Categories, as hypothetical entities in science education, fall under the umbrella term of "concept". These entities are products of reasoning with theoretical inputs provided by formal education [34]. We understand explanations as a vehicle for triggering learning and expressing scientific reasoning that emerges when putting the ideas into a material form of communication (see the next section). Therefore, we interpret students' drawings from the lenses of sociomateriality, both as processes that elicit reasoning and as outcomes of expressed scientific reasoning about a phenomenon that appears to be of high risk.

1.2. Explanations and Students' Scientific Reasoning

Children at school learn about the underlying principles of phenomena and causal relationships, usually but not exclusively in science education. These learning processes are crucial to developing scientific thinking, which is applying the methods or principles of scientific inquiry to reasoning or problem-solving situations [35]. We understand scientific reasoning from a multiple component skills perspective [36], including hypothesizing, experimenting, and evaluating evidence (inferencing, evaluating data, and drawing valid conclusions) [37]. Generating valid conclusions in inquiry processes usually requires explanations. Explanations are particularly characteristic of everyday causal understanding appearing during early childhood [16].

This article studies a specific component, causal scientific reasoning expressed or demonstrated in children's explanations [16] if we take them as a process of intentional knowledge-seeking [36]. Causal scientific reasoning emerges when they need to explain why a specific phenomenon occurs. Constructing explanations requires diverse causal connections [38], which means identifying particular circumstances that can trigger consequences to understand why observed changes or phenomena have a place under certain conditions. Explanations in science education involve scientific knowledge, and they can be based on theory, evidence, and mixed with daily life experiences. Children's scientific reasoning reconciles different kinds of causal explanations about phenomena, such as scientific, natural, and supernatural [17].

Explanations in science education frequently involve abstract knowledge or concepts (i.e., explaining phenomena at an atomic or molecular level mediated by energy transfers). Into a framework for modeling competence, explanations in science classrooms trigger children's abductive reasoning, which is the theory-based attempt of explaining a phenomenon by a cause [38]. Abduction means generating a cause as the best explanation for a phenomenon based on theoretical knowledge [39].

Considering scientific reasoning components, children's use of information to make causal inferences is a complex cognitive task [35,40]. However, this does not imply that young learners cannot express causal reasoning about their natural environment [41]. Wang et al. [42] observed how children between 2 and 5 years old faced causal tasks related to the weight of objects and concluded that, even before primary school, children use causal reasoning in natural environments, although some age-dependent variations were found. Mayer and collaborators [20] measured four scientific reasoning dimensions in everyday situations, one of those was understanding theories. They worked with 155 fourth-grade students in a paper and pencil instrument test. The results showed that children developed their performance in the measured dimensions.

In terms of searching for explanations to make sense of a phenomenon, scientific reasoning is related to the construction of models. A model used for teaching and learning concepts serves as a medium for communication, describing, and explaining [39]. Perkins and Grotzer [40] proposed a selection of causal models based on the level of reasoning sophistication: (a) mechanism, where students can use their experience to make generalizations not always aligned with mechanistic reasoning, moving to more complex and accurate explanations; (b) interaction pattern, a dimension where students use different paths to connect causes and effects; (c) probability, referred to as what could happen; (d) agency, for example when students identify the presence of an agent involved in direct action. Within each of these dimensions, the authors note sublevels of complexity. Based on Perkins and Grotzer's framework and other research studies of causal reasoning in science education, Moreira et al. [9] found that secondary students use complex causal reasonings to develop explanations in a specific chemistry topic. However, their results showed that using mechanistic reasoning does not always guarantee an alignment with scientific theory. Zangori et al. [31] built a rubric based on Perkins and Grotzer's framework [40] and other studies related to reasoning about ecosystems to analyze the causal associations used by third-grade students when they learn about ecosystems. They found the students who had the opportunity to reason using models enhanced their causal reasoning, and intermediate steps towards the use of causal reasoning were identified.

1.3. Scaffolding Explanations in Science Learning

Other studies have developed instructional models or learning progressions to scaffold, assess, and analyze students' explanations at the school level, e.g., [43,44]. These studies have common characteristics; they describe the explanation components and using evidence in their performances. McNeill et al. [43] constructed their instructional model considering Toulmin's framework and standards for science education, describing three explanation components: claim, evidence, and reasoning in the following components:

Claim, an assertion or conclusion that answers the original question; evidence, scientific data that support the claim; the data need to be appropriate and sufficient to support the claim; and reasoning, a justification that links the claim and evidence and shows why the data count as evidence to support the claim by using the appropriate and sufficient scientific principles.

However, a few studies relate explanations and scientific reasoning in evaluative purposes, for instance, highlighting the reasoning expressed by students in their productions. A five-stage comprehensive learning progression of written scientific explanations for the school level was designed by Yao and Guo [44]. At the more basic stages, the students first relate, indirectly, facts and theory through models. When their scientific reasoning evolves, they progressively approach scientifically accepted models. The elements of reasoning appear as a simple causality, moving forward to more complex forms such as probabilistic or correlational reasoning to link the explanations logically [44].

The distinction between school explanations based on evidence versus those based on theory is an ongoing academic discussion. However, we know that the scaffolding process that children need to construct explanations based on their observations, inquiry processes, and evidence is different from the practical support for students to create explanations underpinned by theories, principles, or models that are more abstract entities [43,45]. Among the first group, the studies show that systematically helping students distinguish between the description of the facts, observations, and the emergence of an argument based on evidence is worthy of learning, e.g., [43]. The second group of students' explanations—supported by theories—counts with empirical support of how the use of epistemic tools, such as the Premise–Reasoning–Outcome instructional strategy (P.R.O.) [45] facilitates not only writing of better explanations but enhancing students' cognition and metacognition processes [46]. Thus, in the context of learning to explain phenomena based on theory, we found the research need of a domain-specific instrument to characterize students' reasoning and apply it to explanations.

Previous studies of explanations as a product and process of learning have analyzed verbal or written answers separately, i.e., [9,47,48]. However, this type of analysis has insufficiently captured the complexity and advancement of children's reasoning in learning new scientific concepts [47,49].

Consequently, we focus on generated pictorial representations in drawings, a complementary format vital for children's expression and communication that has been less researched in this field [49]. In addition, Park et al. [50] argued that this type of representation contains implicit information that offers an opportunity to analyze students' ideas and concepts. Indeed, analyzing non-linguistic forms of representation is a more inclusive method to approach students with difficulties with verbal/oral expression [24].

The focus of our work is highlighting and approaching children's reasoning about natural phenomena underpinned by theory from a cognitive perspective. We centered the application of this purpose on student-generated drawings as an alternative form of constructing and communicating explanations to make sense of the causes of a natural phenomenon that might affect their lives, specifically earthquakes. We chose the earthquake phenomena because, in Chile, the country in which this study was conducted, earthquakes are a relatively frequent event that children are familiar with, as the country is in a seismic area. Thus, for the participants living in a geological fault zone, this phenomenon might be more quotidian/frequent or, at least not as unfamiliar as other natural phenomena. Nonetheless, the fourth grade is the first formal opportunity in which students start learning the underpinning theory of this phenomenon, known as Tectonic Plates Theory (henceforth, TPT). Moreover, the transmutation of the daily life self-explanations of phenomena towards scientific explanations based on theory begins at the stage this research took place.

Briefly explained, TPT states that layers and plates form the Earth's internal structure in the static model. Plates move in different directions, giving place to continents as we currently know them. The inner movements of the plates occur mainly in three forms, convergent where plates move towards each other, divergent where plates move away from each other. Lastly, in transform movement, each plate moves sideways compared to the other. As a result of such movements, energy builds up, released through earthquakes, tsunamis, and related events. Therefore, TPT describes movements of plates, explaining the origin and mechanism of earthquakes [51].

We started from the assumption that supporting students in constructing explanations is a high-leverage practice in education [3], implying the development of reasoning processes and more authentic scientific practices in this study regarding TPT.

Our research question was: What characterizes students' expressed reasoning in drawn explanations in the context of learning about earthquakes? The purpose of this article is to shed light on primary students' scientific causal reasoning during a learning sequence at the school, in the context of current challenges in science, as well as to present a novel methodological coding rubric to approach this process. Science education needs to promote students' thinking processes through authentic scientific practices, such as constructing explanations. Thus, this work will contribute to research on primary students' causal reasoning and science education from a cognitive perspective.

2. Materials and Methods

The present study was exploratory with a descriptive and relational scope based on educational practices to inform educational processes. The data set was collected in Spanish and then translated into English by the article's first author for dissemination purposes. The information from the participants was gathered during the science learning sequence about the "Internal Dynamics of the Earth" in 2019. Two stages during the learning sequence were crucial for collecting the data that compose this study, part of a larger project in science education research. These stages are denominated as stage one and stage two, henceforth S1 and S2. S1 represented when the learning unit was started by the teacher, and S2 when the unit finished. It is important to note that this study did not intend to estimate the effectiveness of the teaching unit or determine how the learning opportunities

provided affected students' scientific reasoning skills because the study design did not include an intervention or comparison groups to make those inferences.

Characterizing students' drawings provides opportunities to analyze how instruction and the curriculum need to challenge students' ideas. It is educationally relevant considering that students' and scientific ideas coexist and interplay in their experience of making sense of the world [52]. The instruction helps with a reconstruction of these ideas in the sense of an explanatory coexistence [52].

The learning sequence in our study consisted of approximately 18 h of pedagogical work distributed throughout four weeks. The lessons comprised drawing activities, a group puzzle about Tectonic Plates and watching videos about the consequences of earthquakes, tsunami, and volcano eruptions. The teacher delivered some lectures about Earth Structure and Tectonic Plates' interaction. The students completed learning workbooks about the more dangerous hazards in Chile and socialized a school security plan.

During the learning unit, the learning outcomes were formalized by constructing hand-drawn explanations about the phenomenon of earthquakes. However, the teacher also used other sources to facilitate learning advances regarding tsunamis and volcanic eruptions. The prompt for triggering student drawings used in this study was "Why does the ground move (in a seismic context)? Please draw your explanation in this blank sheet". The instruments and steps of this study were approved by the Pontificia Universidad Católica de Chile's Ethics Committee code number 180514006.

2.1. Participants and Paradigm

The participants were 22 fourth-grade students from families of middle socioeconomic status. The school was selected through purposive sampling and was in an area of Chile identified as being at risk for disaster if an earthquake occurs, near the San Ramon geological fault line in Santiago, the Chilean capital. The partnership with the teacher for the educational purposes of this research included the collaborative design of a learning sequence to help students reason about the causes and consequences of Earth phenomena and, therefore, to construct scientific explanations through drawings. This decision was founded on the participatory research paradigm [53], in which the communities of research are part of the analytic process and the decision-makers in the study.

Although the whole class that composed the group participated in the learning activities, only 22 of the students had parental authorization and their consent to use the drawings for research purposes. Moreover, one student did not attend school the day the teacher allocated time for drawing in S2, and he did not want to do it later. Thus, the final data set consisted of 22 illustrations in S1 and 21 in S2, and some results are presented as percentages.

2.2. Data Analysis and Processing

Our data processing was carried out in three different steps. First, we developed an instrument to categorize the scientific reasoning expressed through drawn explanations following the study by Park et al. [50] about pictorial representations. Then, we used the constructed instrument to analyze a group of students' drawn explanations of earthquakes based on Tectonic Plates Theory (TPT). In the following paragraphs, we describe these two steps.

1. First, we developed an instrument to categorize the scientific reasoning expressed through drawn explanations following the study by Park et al. [50] about pictorial representations when qualitatively learning physics. Their work established three main levels for students' expression: sensory that includes what students sense; unseen substance level, which provides for concrete substances that cannot be seen; and lastly, unseen non-substance that contains those representations about non-concrete and unseen aspects. This prior work was developed with talented students, representing a novel contribution to the field with a limited scope of applicability.

A panel of three experts, including teachers and cognitive psychologists, checked that this first version of the instrument was conceptually adequate, and the levels proposed would be observable in regular primary students learning samples.

2. To expand the applicability and address explanations of regular primary students, we developed a first pilot qualitative analysis of a set of learning samples composed of drawings using the constant comparative method as the primary coding process of Grounded Theory [54]. We created groups of similar drawings and contrasted their main features, discussing the expressed reasoning that could be identified. Then, we went through three flows of activity of the constant comparative method to adjust the instrument to the data: data reduction, data display, and conclusion verification. We also followed the indications by Tang et al. [55] for interpreting specific aspects of children's drawing, such as types of lines for representing movement. Once we went through three rounds of discussion between the authors of this study, clarifications on the instrument were added. We modified the first version of the rubric by adapting the sensory level, the unseen substance level, and the unseen non-substance level of Park et al.'s framework [50], with specific emphasis on explanations of earthquakes based on TPT and an interpretation of the younger student's context-related scientific reasoning.

3. We conducted a qualitative analysis of students' explanations by three independent researchers—also authors of this article—all trained to code the drawings in a blind review process using the instrument developed in the previous steps. The final version of the rubric, which served as a coding framework, is presented in Table 1. The coding process was performed by each researcher independently; a total of 30% of the students' drawings were coded and compared among the three researchers in two rounds. The first round comprised 15% of the data, and the inter-rater reliability was 62%. After discussing the cross-cutting drawings, examples were selected to represent each level (see details in Section 3.3). The disagreements were discussed until a consensus was reached between the three researchers. The second round included a second set of drawings that comprised another 15% of the data set; the inter-coder agreement was 91%, which was considered a high measure of transparency for instrument implementation [56]. The remaining data were coded by one of the researchers considering the high level of prior agreement. The drawings were coded according to the three rows of the rubric. The first identified the main characteristics of the explanations represented in the students' drawings, looking for causes or consequences of the characterized phenomena. The second one centered the attention on the specific elements or details found in the representations. The third one interpreted the type of reasoning the student expressed in each drawing.

Table 1. Coding rubric for primary students' drawn explanations inspired by Park et al. [50].

	Level 0	Level 1	Level 2	Level 3
Description	It is not possible to interpret an explanation connected with the phenomenon from the pictorial representation.	The drawing represents elements within the child's sensory plane, generally as effects of earthquakes, such as the ground's surface movements or movement effects. The information in the representation was not enough to interpret an explanation beyond the child's perceptible plane.	Some elements are beyond the immediate child's sensory or perceptual plane. The drawings present changing aspects, for instance, beneath the ground or views from outside planet Earth. However, it is not evident that these changing entities are related to the interactive basis of TPT, such as movement, friction or a crash of plates, or the dynamics of the internal structure of the Earth.	The drawings include interacting elements that are outside or beyond the child's sensory or perceptual plane (i.e., changing position or moving entities), expressed as a causal explanation of the earthquake, directly connected with TPT (i.e., movement, friction, or a crash between plates, or the dynamics of the internal structure of the Earth). Conceptual inaccuracies are expected even in this level of representation.
Details	Some students wrote "I don't know", drawing a non-related phenomenon from an external observer's view, leaving the paper blank, or presenting incomprehensible elements.	These drawings frequently have a baseline to delimit the ground line (continuum, backstitch, oblique) or function as object support. Some graphics also wrote words related to "movement" or "seism", etc., while others designed zigzags or wavy lines to represent the consequences of movement on the objects.	These drawings commonly represent a baseline to express a division between the elements perceived and the not perceived but conceptualized and represented as the possible causes of earthquakes. This conceptualization attempts to express a causal relationship between the consequences of the earthquake and its origin.	The drawings include the causes and consequences of the phenomenon, usually with arrows or labels indicating the name of the components (i.e., epicenter, interaction, etc.) or the direction of the movement. These drawings are precursor models used to express a causal relationship between the phenomena and the underpinning theory.

Table 1. Cont.

	Level 0	Level 1	Level 2	Level 3
Reasoning (interpreted)	Students' expressed reasoning is not possible to be interpreted from these types of drawings.	It is a sensory level of reasoning because the cognitive operation is based on entities or elements within the students' perception of their senses.	The reasoning includes elements or processes beyond the sensory experience, attempting to express causality, nonetheless, not yet at a level that uses the parts of a theory to represent causal processes or ongoing mechanisms.	There is a qualitative leap of children's reasoning towards thinking with non-visible theories or non-perceived elements to explain processes or ongoing mechanisms as the cause of phenomena, using theories, abstract concepts, or models. Thus, reasoning at this level is at a more sophisticated stage than in the previous levels.

3. Results

This section describes first the coding framework and the rubric developed to characterize the students' expressed reasoning through drawn explanations. Secondly, we present the application results for fourth graders' drawings based on the main elements that constituted the participants' explanations based on theory. After this, we show the main trends of this group of participants' reasoning levels coded at the beginning and the end of a learning unit in context-based science learning related to earthquakes to illustrate a practical application of this novel approach. These results are presented as an example of the possible analysis of drawn explanations using the developed instrument but do not limit the application to one phenomenon only. Finally, we illustrate the composition of each reasoning level with some drawing examples, highlighting their inferior and superior anchor to orient teachers and researchers on the transitions from one reasoning level to the next one in the case of TPT.

3.1. Instrument for Characterizing Scientific Reasoning in Drawings

The instrument developed in our study takes the form of a comprehensive rubric which works as a coding system to facilitate the assignment of levels, and the characterization of primary students' expressed reasoning through drawn explanations. The rubric allows a description of both the characteristics of the domain-specific drawings and the reasoning level that might be externally interpreted.

Precisely, the rubric developed in this research (Table 1) consists of a three-level grid oriented to progressively identify levels of scientific reasoning in primary students, which are presented as columns. However, the first column represents a level 0 for drawings under the category of missing. As the instrument was applied to learning about earthquakes, its specification for Earth Science phenomena and TPT theory is included. We decided to base our work on distinguishing between perceptual planes expressed as input for interpreting reasoning and the connection between the explained phenomenon and its underpinning theory. This decision sought broader use of this approach to characterize early stages of students' scientific learning based on theories for modeling and explaining phenomena. In the topic of this study application, this stage corresponds to the fourth grade.

Additionally, the instrument added a minimum level used to code the learning samples that could not be categorized or did not answer the cognitive demand of the task, which is quite frequent in young children or during initial learning processes. We expected that students' drawn explanations move throughout the starting levels, from concrete or straightforward stages—based on their previous experiences, highlighting a sensorial focus—to more abstract ideas considering causal links, likewise expressing more complex reasoning. Furthermore, the rubric would make visible the sophistication of the students' expressed reasoning and understanding of the underpinning theory. Thus, the levels proposed in our instrument could also be used as an emergent learning progression.

The three rows of the rubric present elements as follows.

1. The first one describes the main characteristics of the explanations represented in the students' drawings, emphasizing the differentiation between their expressed sensory plane and the connection with the theory.

2. The second row presents the specific elements or details found in fourth-graders learning about a particular phenomenon, in this case, earthquakes, as an application of the first row to domain-specific learning samples.

The two first parts of the instruments may be adapted for working with other theories or phenomena.

3. The third row describes the interpreted scientific reasoning in connection with the sensory planes, the causality, and the usage of theory as a more abstract step in the students' cognitive processes when learning science. This part of the instrument is not associated with a singular theory; thus, it does not need adaptation to apply other topics.

The interpretation of reasoning is suitable to be used by educators or researchers in other learning topics or areas beyond Earth Science when students construct explanations based on scientific theories. It constitutes the first contribution of our work related to science learning research transcending the specific theory and expanding the cognitive process of causal reasoning rather than focusing on the learning accuracy of scientific concepts.

3.2. Trends in the Participants' Reasoning Levels

Considering the categorization results of the participants' explanations using the instrument described earlier (Figure 1), we observe that in the early stage in the formal process of learning—called Stage 1—(S1) before the learning unit began at the school, 28% of the students' explanations did not achieve the minimum level for categorization. Consequently, level 0 was assigned, as shown in Figure 1. In comparison, 24% of the drawings were categorized at level 1 for reasoning and 48% at level 2. This result means that most fourth-grade students could express reasoning about earthquakes with attempts to go beyond their immediate perception plane, representing elements that might constitute a causal explanation later, even with no formal instruction. However, none of the drawings reached level 3, causal reasoning based on aspects of TPT. Thus, we observed that some of them might have had an intermediate level of reasoning even with no formal instruction in this group of students. Nonetheless, establishing connections between the phenomenon and the theory in the form of a causal explanation in the drawings was difficult for the students.

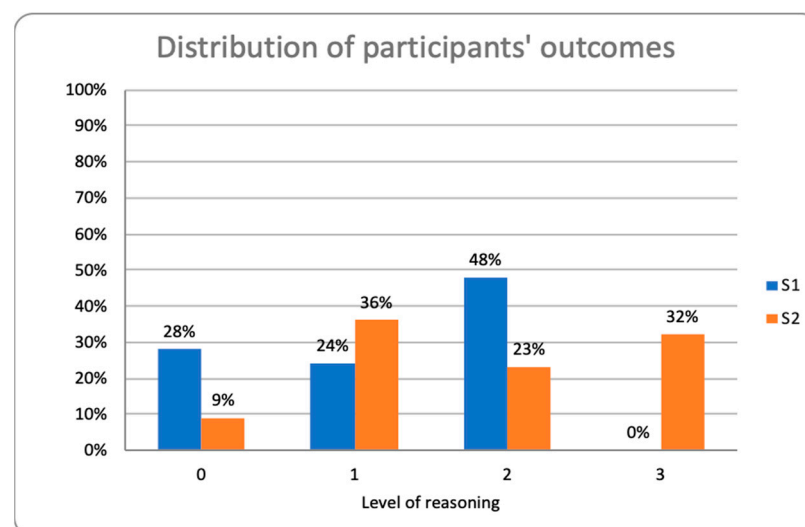


Figure 1. Participants' reasoning levels at the beginning and end of the unit.

In Stage 2—(S2), after the learning unit about the internal dynamics of the Earth was implemented, we saw a reduction in the percentage of drawings at level 0, with only 9% of the students' samples categorized as such. There was a proportional increase in level

1 illustration, with 36% classified as level 1 instead of the 24% obtained at S1. It is interesting to note that student representations categorized as level 2 decreased from 48% to 23% compared to S1; however, this fact is attributed to an increase in the drawings categorized in level 3, comprising 32% of the total. Thus, we conclude that, after participating in a formal learning process about earthquakes, it is likely that most of the participants in this group of students could express more sophisticated reasoning and a causal link in their drawn explanations. Nonetheless, 45% of children did not show cognitive operations with unobservable entities or logically connect the causes and consequences of earthquakes, even after the learning unit was finished.

3.3. Characterization and Examples of Reasoning Levels Interpreted from Drawings

This subsection presents descriptions, main features, and examples for each level identified, representing the finest-grain analysis of student drawings. It is worth remembering that, in the context of learning about Earth Sciences, the task demanded was “draw or represent here your explanation about Why does the ground moves?”

Level 0: It is impossible to interpret an explanation connected with the earthquake phenomenon from the pictorial representation. For instance, some students wrote “I don’t know”, drew a non-related phenomenon from an external observer’s view, left the paper blank, or presented elements that were incomprehensible for the researchers in the light of the question demanded by the task. Thus, we could not interpret the students’ expressed reasoning from these types of drawings. The authors of this work considered this level as missing data. This means that interpretable reasoning could not be obtained from an external viewer solely from drawings regarding the question given. However, other researchers might combine these types of illustrations with oral or written explanations; thus, the character of missing data would change. Some examples of pictorial representations categorized in this level in the current study are presented in Figure 2.



Figure 2. Examples of level 0. Drawing (A) shows a volcano, (B) represents the Earth planet and where Chile is.

Level 1: The student drawing represents elements within their sensory plane, generally as effects or consequences of the earthquake phenomenon, recognizable as movements of the ground’s surface or results of the movement. The information derived from the representation was insufficient for the researchers to interpret an explanation beyond the child’s perceptible plane, for instance, based on non-visible entities. These drawings frequently have a baseline to delimit the ground line (in a continuum, backstitch, or oblique) or function as object support. Some graphics also wrote words related to “movement” or “seism”, etc., while others designed zigzag or wavy lines to represent the consequences of movement on the objects, as Figure 3 shows. Thus, we interpreted these drawings as a sensory level of reasoning because the cognitive operation is based on entities or elements within the students’ perception of their senses.



Figure 3. Examples of level 1. Drawing (A) illustrates a field with plants moving, (B) a ground line with scared children moving, and a happy face below the baseline.

Level 2: Some representations or elements are beyond the students' primary sensory or perceptual level. The drawings in this category (Figure 4, in which we have translated what the students wrote in their drawings) usually present changing elements, for instance, beneath the ground, or views from outside planet Earth, commonly represented by a baseline–ground line or object support–to express a division between the elements perceived by children and the elements not perceived but conceptualized and represented as the possible causes of earthquakes. In these types of drawings, we observed an attempt at expressing a causal relationship between the consequences of the earthquake (i.e., beyond the baseline) and their origin (i.e., beneath the baseline); however, it is not evident that these changing entities are related to the interactive basis of TPT, such as movement, friction or a crash of plates, or the dynamics of the internal structure of the Earth. Thus, we interpret a more complex level of reasoning than in level 1 because children are reasoning through elements or processes that are further from their immediate sensory experience and trying to express causal thinking, nonetheless not yet at a level that uses the parts of the theory to represent a causal process or ongoing mechanism.



Figure 4. Examples of level 2. Drawing (A) represents moving buildings on the surface and Earth layers beneath, (B) shows a broken building and elements under the base line labeled as “Plates”.

Level 3: The representations in this level were more complex in comparison with level 2. The drawings include elements outside or beyond the children's primary sensory or perceptual level. However, the difference with level 2 is that, in level 3, these components are interacting, changing position, or moving. These concepts are expressed as a causal explanation of the earthquake, directly connected with TPT, such as movement, friction or a crash between plates, or the Earth's internal structure dynamics. We observed drawings that included the causes and consequences of the phenomenon, usually with arrows or labels indicating the name of the components (i.e., epicenter, interaction, etc., illustrated in Figure 5) or the direction of the movement. Thus, we interpret these drawings as precursor models used by the participants to express a causal relationship between the phenomena and the underpinning theory, which means a qualitative leap of children's reasoning towards thinking with non-visible theories to explain processes or ongoing mechanisms. It is worth noting that we made no judgment of the conceptual accuracy presented through the representation.

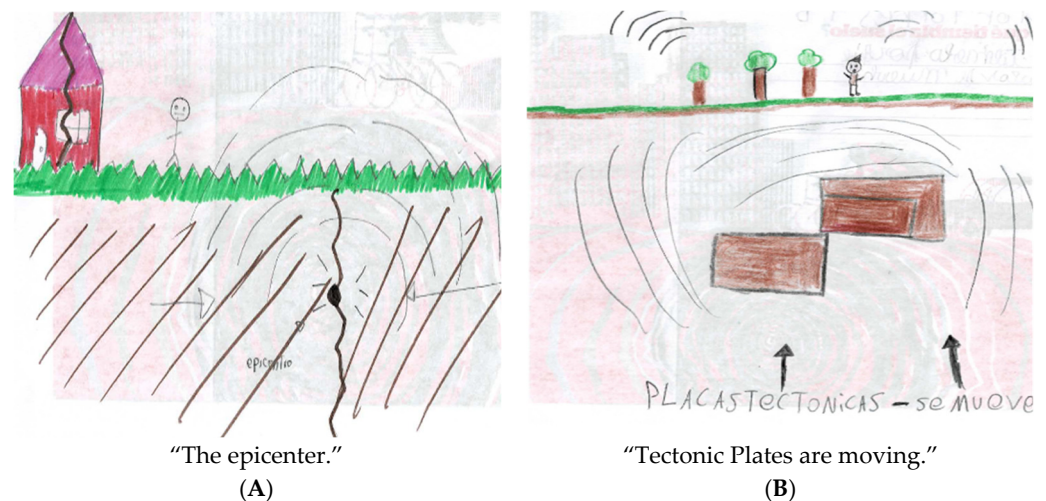


Figure 5. Examples of level 3. (A) represents a damaged house, a sad person on the surface and under the herb line a point of interaction labelled "the epicenter" with facing arrows. (B) shows trees moving, a scared person above the baseline and two blocks moving labelled as "Tectonic plates are moving" under the baseline.

Conceptual accuracy refers to the degree of content correctness in the scientific use of concepts, terms, or postulates in the drawing. Although in other works with secondary students' explanation, a conceptual inaccuracy in the written explanation implies coding in level 0, e.g., [44], in this study, we consider that primary students can have inaccuracies expected because they had only started to learn about the content. Thus, we decided to give value even to explanations that were not totally precise but showed the advance in the reasoning process. For instance, in Figure 5, student's drawing A represented the causes of earthquakes under the baseline, reasoning with abstract entities, represented a model of interaction, signaling a black point where the energy releases as "the epicenter." Although the correct term should be "the hypocenter," we made no judgment of the conceptual accuracy in the representation and consider it is an advance in the expressed reasoning regarding levels 1 or 2. Thus, we categorized it at level 3.

3.4. Boundaries for Interpreting "Qualitative Leaps" to a Superior Level

Our research found three qualitative leaps of expressed reasoning in students through drawn explanations, which help us interpret a hypothetical progression of reasoning. The first one (a) marks the level at which we can affirm interpretable reasoning about the phenomenon. The second one (b) refers to an advance from the upper anchor of level 1 to

the inferior anchor of level 2. The third leap (c) occurs between the upper anchor of level 2 to the low anchor of level 3.

- (a) The entry point to the hypothetical progression of reasoning is the connection of the explanation with the phenomenon of interest. In this case, we observed the leap between level 0 and level 1 when the students represented the effects or consequences of earthquakes. In addition, they recognized that, in the context of learning about the internal dynamics of the Earth, the cognitive task that required drawing “why does the ground move?” involves a specific phenomenon—an earthquake. Level 1 is minor complex because the student only needs to identify a logical connection within the task’s context. For example, in Figure 2B, the planet Earth drawing was categorized at level 0, missing data. However, in Figure 3A, at the bottom anchor of level 1, we considered the black lines around the plants in the soil to represent movement, according to the categories by Tang et al. [55], which signal a consequence of the earthquake.
- (b) Comparing the upper anchor of level 1 to the inferior anchor of level 2, we can observe the qualitative leap that focuses beneath the ground level as a baseline. In Figure 3B, even though there is a line that might divide the perceptual plane from the non-perceptual plane, beneath this line, there are no recognizable elements. On the contrary, in Figure 4A, it is possible to observe the same ground line but with a representation of the Earth’s layers similar to the static model. Thus, we interpreted the increased complexity of the child’s recognition of possible causes of the phenomenon with an incipient link to the TPT.
- (c) Between the upper anchor of level 2 and the inferior anchor of level 3, we interpret a leap signaled by some representation elements connecting with the modeling process in science education, in the labeling of “Plates” in Figure 4B. However, no interaction between the components was expressed. The sophistication was demonstrated by the more explicit representation of the interaction between unobservable entities. Figure 5A,B represents cause, consequences, and activities between the components of TPT. They show reasoning with theory to explain a natural phenomenon.

4. Discussion

In this study, we sought to explore the characteristics of students’ expressed reasoning through drawn explanations in the context of learning about earthquakes at an early stage of formal instruction. We developed an instrument based on previous research to elicit and analyze fourth graders’ scientific reasoning based on theory through their drawn explanations. The analysis allowed the recognition of three levels of scientific reasoning, which were possible to characterize in the participants of this study. Consequently, our findings answer the question proposed: What characterizes students’ expressed reasoning in drawn explanations in the context of learning about earthquakes?

In summary, at level 0, topic-specific reasoning was not interpretable from the representational explanation. In contrast, at level 1, students’ reasoning was based mainly on the perceptible entities associated with the consequences of the phenomena. Drawings characterized as level 2 showed that children’s reasoning starts to connect some theory elements as a first attempt to explain the causes of a phenomenon. Despite this, levels 1 and 2 lack causal relations using the theory. Finally, in level 3, students could express their scientific reasoning about the phenomenon by linking elements of TPT to explain the causes and effects of a phenomenon as a precursor model, considered as cognitive schemata compatible with scientifically appropriate knowledge [27].

Moreover, our study found qualitative leaps between the children’s levels of expressed scientific reasoning focused on the connection with the phenomena under investigation, the emergence of the divisions of the perceptual and non-perceptual plane, and the presence of recognizable elements of the theory as part of the representation of the explanation. Some of the more advanced features expressed by the participants in our study presented similar characteristics to those of Perkins and Grotzer [40]. Specifically, we interpreted sophistica-

tion from static comprehension to an interactive activity between the non-observable or theoretical entities.

According to Yao and Guo [44], the students first relate indirectly to facts and theory through models before their scientific reasoning evolves, progressively approaching scientifically accepted models. Analyzing students' drawings as an expression of their reasoning process gave us evidence for interpreting more sophisticated reasoning during the learning unit and students' drawings as precursors of scientific models. This idea might be construed from a transformation of embedded intuitive theories through language [18] and deliberate thinking [19]. Moreover, our study expanded the literature to other forms of capturing advances of students' reasoning through their creative activity of drawing explanations, which represents a complement to the current instruments to analyze students' written explanations [31,43,44].

However, simultaneously analyzing and fostering students' explanation construction based on theories, principles, or concepts is still a challenge at the early stages of formal learning [6]. Given this, we need to understand that students are still constructing the meaning of the scientific concepts involved when explaining. In primary education, they also develop essential skills such as explaining for scientific purposes or using models to explain the world. Thus, we emphasize the importance of supporting students to build these capacities and not underestimating their possible ability to express their scientific reasoning and knowledge through formats more familiar to them, such as drawings. Combining forms for approaching scientific reasoning and learning might mean a synergistic effort to scaffold the emergence and sophistication of reasoning, the conceptual understanding of children, and the development of essential skills. Our results resonate with prior research showing the need to combine diverse data sources to interpret children's scientific learning [27].

Park et al. [50] discussed pictorial representation as a complementary format to explore students' ideas. In this, they argued that drawings involve implicit information that is connected to other external representations. Indeed, the ways students express themselves about a concept or idea might be different when they do it verbally and pictorially, or exclusively verbally. We believe that for younger students, it is through drawings or representations that they are building scientific ideas and connecting them to other types of representations. We know that for students to construct scientific concepts, multimodal languages support processes related to sensemaking, scientific explanation construction, and scientific concept development [21]. It implies that employing exclusively visual or verbal representation during teaching might limit students' learning process. Considering pictorial representations as part of multimodal language supports students in building concepts that are vehicles for expressing their reasoning. By having students use verbal communication only for concept construction, incorporating pictorial representations might result in more prosperous, more robust, and connected ideas for concepts construction, perhaps involving a re-conceptualization due to changing modalities. This is because constructing explanations seen from a sociocultural perspective is a knowledge integration learning artifact, in which the students connect what is already known with their experiences and conceptual elements to give scientific support for certain phenomena [25].

The instrument used to analyze scientific reasoning based on theory for primary students was demonstrated to be sensitive enough to detect the sophistication of these elements of reasoning during a learning unit of eighteen hours in the context of this study. Specifically, we observed an increase in level 1 and 3 categorized drawings between the learning sequence's beginning—S1—and the end—S2 (Figure 1). Thus, we can conclude that, after participating in a formal instruction process, some participants in our study could express more sophisticated reasoning with a causal link in their drawn explanations. We agree that explaining phenomena provides an optimal scenario to connect students with socio-scientific issues [30,31], and our study adds that student drawings can be a source of expressed reasoning and, at the same time, a learning activity that activates and allows enacting or triggering of specific systems of reasoning.

Nonetheless, the instrument allowed the identification of a significant group of participants who did not show evidence of operating cognitively with unobservable entities to connect the causes and consequences of the phenomena under study. After the learning unit was finished, this gap was observed, with students immersed in a high-risk context, adding familiarity with the phenomenon. We recommend providing opportunities to learn to link phenomena and their causes in this and several other topics and conducting more research to determine the obstacles to student advancement in reasoning levels. Still, we observed instances of expressed reasoning regarding context-related situations before formal learning started at school. The entry point to the hypothetical reasoning progression was the connection of the explanation with the studied phenomenon. This finding coincides with studies that show the starting point for explanation-construction is the phenomenon [4], which helps to afford the need to generate explanations. By fourth grade, Chilean students have likely already had some daily life experiences with earthquakes and can nurture their reasoning process about the environment in which they live. Thus, the fact that our study considered the early stages of formal learning and identified what ideas the students had already formed in their representations for constructing explanations is valuable. Further research could illuminate the role of local context in early scientific reasoning levels, not only on how scientific reasoning about earthquakes develops throughout the school trajectory but also extending the use of such instruments to other subjects, areas, or demanding tasks.

This study has some limitations. First, using a strategy designed in a different context and language might cause cross-cultural issues. We adapted the frame suggested by Park et al. [50] according to the context of the study. Still, we also acknowledge the particularities of Chile as having a high risk of disaster (e.g., earthquakes). Thus, the learning approach to these phenomena may vary from those whose context does not include risk or whose geographical reality is very different. However, this point also represents a possible subject for future researchers to explore: the extent to which proximity to a phenomenon might imply a variation in the way students think about it.

Additionally, some elements of students' drawn explanations went beyond the frames of our analysis, for instance perspectives from outside the planet that combined astronomical concepts. Although we treated those features as exceptions in our study, perhaps representing a limitation, we believe a second perspective on these types of data is crucial to challenge adults' beliefs about the abstraction capacity of children and the way they visualize phenomena and their causes. Moreover, we recognize our study has a small sample size for going beyond descriptive analysis. Thus, we encourage further research to work with larger groups of students for complementary validation purposes.

Regarding the validation of the rubric, in this study, we went through a content validation through a panel of experts and a small pilot study before analyzing the data sets. Due to the small sample size and the study's exploratory nature, we could not run factor analysis or more sophisticated processes, strengthening the significance and or generalizability of the results.

Nonetheless, we consider this study as a first approach interpreting primary students' reasoning in phenomena explained by theory, with an educational significance in the field of science education. Other researchers might take the advances of our work and, for instance, compare pre–post drawings in specific groups of students, or use a repeated-measurements design focusing on learning the topic or conceptualization of the phenomena. Hence, we suggest future research gathering evidence of the leaps shown in our study but exploring them in the light of learning progressions of individual students. This exploration might complement the current results to emphasize the connections between understanding phenomena, theories, or concepts and learning, to establish learning trajectories in science education.

5. Conclusions

The current study allowed us to characterize students' scientific reasoning through drawn explanations. We presented a helpful instrument to identify cognitive leaps between concrete expressed reasoning levels and more abstract ones, including causal links between phenomena and theory. It is a methodological innovation to approach young students' learning and reasoning development from an interdisciplinary perspective that combines education and cognitive science. Our research explicitly links science learning and cognition by highlighting and approaching children's reasoning about natural phenomena underpinned by theory. This development expands the current instruments available to notice the complexity of scientific reasoning of young children when they are at the first moments of learning models, theories, or abstract postulates that sustain the causes of phenomena. There is a methodological advance considering that most of the current instruments relate to explanations based on evidence in written formats.

In applying the developed rubric, we observed sophistication in students' scientific reasoning when provided a formal learning opportunity, resulting in some students progressively connecting their ideas to a scientific theory. Our study allowed exploration of students' progressive development of the causal reasoning required to construct explanations. Constructing explanations based on theory from primary school is a relevant teaching and learning practice to develop at an early stage of learning, considering that secondary and college students have limitations to using their scientific knowledge to establish causal links when they construct explanations. Furthermore, identifying scientific reasoning levels at the early stages of learning allows conceptualization of scientific reasoning as a trajectory. Thus, we can observe more precisely where students begin this form of complex knowledge and how it will eventually progress. By identifying and understanding this trajectory and the qualitative leaps, teachers, educators, and researchers can better scaffold the learning process and the development of context-related scientific reasoning, providing opportunities to support this development promptly. The detailed description of these findings helps researchers interested in this field adapt, reframe, and test in different ways the analysis we have done, allowing projection of transference of the interpreted reasoning of the rubric of this research to other topics. It would make the progressive approach to thinking in different disciplines visible and promote students' reasoning in the school. This idea resonates with theoretical frameworks used for understanding of the construction of explanations as epistemic processes, which broadens the interest of this article to other areas beyond the content of the application in our study.

Teachers' support of children's reasoning in the classroom might take the form of distributed scaffolding. For instance, giving prompts with initial questions such as in the present study "why do you think this phenomenon happens", and moving forward to students to revise and enrich their initial explanations during the learning of the content advance. The scaffolding seeks to transfer the responsibility gradually to the student, promoting autonomy. In primary education, where students are diverse in autonomy degrees, generating group discussions about explanations is an option, considering that science practices also imply peer-reviewing ideas and claims to compare and contrast to evaluate their scope and limitations. This strategy also connects with positioning science construction as a collective activity, introducing children to elements of Nature of Science. We strongly believe that classroom activities oriented to develop students' reasoning processes should encourage students to express their ideas in diverse formats, such as the causes of phenomena. Then, linking those with the scientific support through concepts, theories, or postulates that are usually more abstract entities to reason. However, this approach needs teachers to consider that students' common sense is part of their implicit theories that allow them to make sense of emergent phenomena, thus relevant for transformation and not represent merely knowledge to discard during science lessons. We know that teachers tend to suppress ideas that might look wrong as they are expressed in more traditional science classrooms. Still, we want to stress that responsive science teaching gives value to the students' existing ways of thinking to construct new understanding, further develop

their reasoning into a more scientific one, managing the supports strategically that students need promptly.

Furthermore, our work supports understanding primary students' reasoning considering current educational challenges, affording students' thinking processes through authentic practices, such as constructing explanations based on context-related phenomena. Moreover, we see the explanation-construction of relevant phenomena as a participatory action for responsible citizenship that can be implemented in primary education to promote high-leverage practices such as explaining and modeling, as was mentioned in our theoretical framework. Thus, we highlight that, even at the early stages of formal science learning, students can transform their ideas into expressions of context-related reasoning, for instance, through drawings that act as learning samples of explanations represented at the first stages of their models to explain natural phenomena. This fact emphasizes the importance of recognizing young children as active constructors of knowledge, showing that some can go beyond their immediate experience to logically link a phenomenon with its underpinning theory. Constructing explanations about world phenomena and expressing students' reasoning in formats aligned with their action, drawing creative activity is a more abstract and complex process worthy of considering by researchers, educators, and teachers interested in the multidisciplinary innovations for understanding learning processes and outcomes.

Author Contributions: Conceptualization, V.M.C. and P.M.M.; methodology, V.M.C. and P.M.M.; formal analysis, V.M.C., P.M.M. and P.G.M.; writing—original draft preparation, V.M.C. and P.M.M.; writing—review and editing, P.G.M.; project administration, V.M.C.; funding acquisition, V.M.C. All authors have read and agreed to the published version of the manuscript.

Funding: CONICYT/FONDECYT/11181050, currently Agencia Nacional de Investigación y Desarrollo ANID/FONDECYT/ 11181050.

Institutional Review Board Statement: This study was approved by the ethics committee of the Pontificia Universidad Católica de Chile, under number 180514006.

Informed Consent Statement: All participating children had active parental consent, meaning that parents were informed and agreed to their child's participation in the study.

Data Availability Statement: Data are available if required.

Acknowledgments: We thank the article's anonymous reviewers for their thoughtful comments.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the study's design, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

- O'Flaherty, J.; Beal, E.M. Core competencies and high leverage practices of the beginning teacher: A synthesis of the literature. *J. Educ. Teach.* **2018**, *44*, 461–478. [CrossRef]
- McNeill, K.L.; Krajcik, J. Scientific explanations: Characterizing and evaluating the effects of teachers' instructional practices on student learning. *J. Res. Sci. Teach.* **2008**, *45*, 53–78. [CrossRef]
- Braaten, M.; Windschitl, M. Working toward a stronger conceptualization of scientific explanation for science education. *Sci. Educ.* **2011**, *95*, 639–669. [CrossRef]
- Papadouris, N.; Vokos, S.; Constantinou, C.P. The pursuit of a "better" explanation as an organizing framework for science teaching and learning. *Sci. Educ.* **2017**, *102*, 219–237. [CrossRef]
- Forbes, C.; Lange, K.; Möller, K.; Biggers, M.; Laux, M.; Zangori, L. Explanation-Construction in Fourth-Grade Classrooms in Germany and the USA: A cross-national comparative video study. *Int. J. Sci. Educ.* **2014**, *36*, 2367–2390. [CrossRef]
- Cabello, V.M.; Sommer, M. Andamios de retiro gradual. Parte 1: Visibilización del pensamiento en la construcción de explicaciones científicas escolares. *Estud. Pedagógicos* **2020**, *46*, 257–267. [CrossRef]
- Meneses, A.; Hugo, E.; Montenegro, M.; Valenzuela y Ruiz, M. Explicaciones científicas: Propuestas para la enseñanza del lenguaje académico. *Boletín De Lingüística* **2018**, *30*, 134–157.
- Hsu, Y.-S.; Lai, T.-L.; Hsu, W.-H. A Design Model of Distributed Scaffolding for Inquiry-Based Learning. *Res. Sci. Educ.* **2015**, *45*, 241–273. [CrossRef]

9. Moreira, P.; Marzabal, A.; Talanquer, V. Using a mechanistic framework to characterise chemistry students' reasoning in written explanations. *Chem. Educ. Res. Pract.* **2019**, *20*, 120–131. [CrossRef]
10. Tang, K.-S.; Putra, G.B.S. Infusing Literacy into an Inquiry Instructional Model to Support Students' Construction of Scientific Explanations. In *Global Developments in Literacy Research for Science Education*, 1st ed.; Tang, K.-S., Danielsson, K., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 281–300. [CrossRef]
11. Wang, C.-Y. Scaffolding middle school students' construction of scientific explanations: Comparing a cognitive versus a metacognitive evaluation approach. *Int. J. Sci. Educ.* **2015**, *37*, 237–271. [CrossRef]
12. Yeo, J.; Gilbert, J.K. Constructing a scientific explanation—A narrative account. *Int. J. Sci. Educ.* **2014**, *36*, 1902–1935. [CrossRef]
13. Figueroa, J.; Meneses, A.; Chandia, E. Academic language and the quality of written arguments and explanations of Chilean 8th graders. *Read. Writ.* **2018**, *31*, 703–723. [CrossRef]
14. Rappa, N.A.; Tang, K.S. Integrating disciplinary-specific genre structure in discourse strategies to support disciplinary literacy. *Linguist. Educ.* **2018**, *43*, 1–12. [CrossRef]
15. Lombrozo, T.; Vasilyeva, N. Causal explanation. In *Oxford Handbook of Causal Reasoning*, 1st ed.; Waldmann, M., Ed.; Oxford University Press: New York, NY, USA, 2017; pp. 415–432.
16. Wellman, H.M.; Liu, D. Causal reasoning as informed by the early development of explanations. In *Causal Learning: Psychology, Philosophy, and Computation*; Gopnik, A., Schulz, L., Eds.; Oxford University Press: Oxford, UK, 2007; pp. 261–279. [CrossRef]
17. Legare, C.H. The contributions of explanation and exploration to children's scientific reasoning. *Child Dev. Perspect.* **2014**, *8*, 101–106. [CrossRef]
18. Gerstenberg, T.; Tenenbaum, J.B. Intuitive theories. In *Oxford Handbook of Causal Reasoning*, 1st ed.; Waldmann, M., Ed.; Oxford University Press: New York, NY, USA, 2017; pp. 515–548.
19. Evans, J. Dual-process theories. In *The Routledge International Handbook of Thinking and Reasoning*, 1st ed.; Ball, L.J., Thompson, V.A., Eds.; Routledge: New York, NY, USA, 2017; pp. 173–188.
20. Mayer, D.; Sodian, B.; Koerber, S.; Schwippert, K. Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learn. Instr.* **2014**, *29*, 43–55. [CrossRef]
21. Mortimer, E.F.; Wertsch, J.V. The architecture and dynamics of intersubjectivity in science classrooms. *Mind Cult. Act.* **2003**, *10*, 230–244. [CrossRef]
22. Berland, L.K.; Reiser, B.J. Making sense of argumentation and explanation. *Sci. Educ.* **2009**, *93*, 26–55. [CrossRef]
23. Oliveira, A.W. Engaging students in guided science inquiry discussions: Elementary teachers' oral strategies. *J. Sci. Teach. Educ.* **2010**, *21*, 747–765. [CrossRef]
24. Williams, M.; Tang, K.S. The implications of the non-linguistic modes of meaning for language learners in science: A review. *Int. J. Sci. Educ.* **2020**, *42*, 1041–1067. [CrossRef]
25. Bell, P.; Linn, M.C. Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *Int. J. Sci. Educ.* **2000**, *22*, 797–817. [CrossRef]
26. Contessa, G. Scientific representation, interpretation, and surrogate reasoning. *Philos. Sci.* **2007**, *74*, 48–68. [CrossRef]
27. Ravanis, K.; Christidou, V.; Hatzinikita, V. Enhancing conceptual change in preschool children's representations of light: A sociocognitive approach. *Res. Sci. Educ.* **2013**, *43*, 2257–2276. [CrossRef]
28. Aleksandrovna Makarova, E.; Lvovna Makarova, E.; Mikhailovna Varaksa, A. Education process visualization in metacognition development and sustainability. *IJCRSEE* **2017**, *5*, 65–74. [CrossRef]
29. Smith, D.C.; Neale, D.C. The construction of subject-matter knowledge in primary science teaching. *Teach. Teach. Educ.* **1989**, *5*, 1–2. [CrossRef]
30. Zembal-Saul, C.; McNeill, K.L.; Hershberger, K. *What's Your Evidence?: Engaging K-5 Children in Constructing Explanations in Science*; Pearson Higher Ed.: Boston, MA, USA, 2013.
31. Zangori, L.; Ke, K.; Sadler, D.; Peel, A. Exploring primary students causal reasoning about ecosystems. *Int. J. Sci. Educ.* **2020**, *42*, 1799–1817. [CrossRef]
32. Driver, R.; Rushworth, P.; Squires, A.; Wood-Robinson, V. *Making Sense of Secondary Science: Research into Children's Ideas*; Routledge: London, UK, 2005.
33. Benedict-Chambers, A.; Aram, R. Tools for teacher noticing: Helping preservice teachers notice and analyze student thinking and scientific practice use. *J. Sci. Teach. Educ.* **2017**, *28*, 294–318. [CrossRef]
34. Tsankov, N. The transversal competence for problem-solving in cognitive learning. *IJCRSEE* **2018**, *6*, 67–82. [CrossRef]
35. Zimmerman, C. The development of scientific thinking skills in elementary and middle school. *Dev. Rev.* **2007**, *27*, 172–223. [CrossRef]
36. Schlatter, E.; Lazonder, A.W.; Molenaar, I.; Janssen, N. Individual Differences in Children's Scientific Reasoning. *Educ. Sci.* **2021**, *11*, 471. [CrossRef]
37. Schiefer, J.; Golle, J.; Tibus, M.; Oschatz, K. Scientific reasoning in elementary school children: Assessment of the inquiry cycle. *J. Adv. Acad.* **2019**, *30*, 144–177. [CrossRef]
38. Rocksén, M. The Many Roles of "Explanation" in Science Education: A Case Study. *Cult. Stud. Sci. Educ.* **2016**, *11*, 837–868. [CrossRef]
39. Upmeier zu Belzen, A.; Engelschalt, P.; Krüger, D. Modeling as Scientific Reasoning—The Role of Abductive Reasoning for Modeling Competence. *Educ. Sci.* **2021**, *11*, 495. [CrossRef]

40. Perkins, D.N.; Grotzer, T.A. Dimensions of causal understanding: The role of complex causal models in students' understanding of science. *Stud. Sci. Educ.* **2005**, *41*, 117–166. [CrossRef]
41. Lawson, A. The nature and development of scientific reasoning: A synthetic view. *Int. J. Sci. Math. Educ.* **2004**, *2*, 307–338. [CrossRef]
42. Wang, Z.; Williamson, R.A.; Meltzoff, A.N. Preschool physics: Using the invisible property of weight in causal reasoning tasks. *PLoS ONE* **2018**, *13*, e0192054. [CrossRef]
43. McNeill, K.; Lizotte, D.; Krajcik, J.; Marx, R. Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *J. Learn. Sci.* **2006**, *15*, 153–191. [CrossRef]
44. Yao, J.X.; Guo, Y.Y. Validity evidence for a learning progression of scientific explanation. *J. Res. Sci. Teach.* **2018**, *55*, 299–317. [CrossRef]
45. Tang, K.-S. Constructing scientific explanations through premise–reasoning–outcome (PRO): An exploratory study to scaffold students in structuring written explanations. *Int. J. Sci. Educ.* **2016**, *38*, 1415–1440. [CrossRef]
46. Tang, K.-S. The use of epistemic tools to facilitate epistemic cognition & metacognition in developing scientific explanation. *Cogn. Instr.* **2020**, *38*, 474–502. [CrossRef]
47. Sommer, M.; Cabello, V.M. Andamios de retiro gradual. Parte 2: Apoyos a la construcción de explicaciones en ciencia primaria. *Estudios Pedagógicos* **2020**, *46*, 269–284. [CrossRef]
48. De Andrade, V.; Freire, S.; Baptista, M. Constructing scientific explanations: A system of analysis for students' explanations. *Res. Sci. Educ.* **2019**, *49*, 787–807. [CrossRef]
49. Panagiotis, E.; Chachlioutaki, M. Reanalysing children's responses on shadow formation: A comparative approach to bodily expressions and verbal discourse. *Int. J. Sci. Educ.* **2017**, *39*, 2508–2527.
50. Park, J.; Chang, J.; Tang, K.S.; Treagust, D.F.; Won, M. Sequential patterns of students' drawing in constructing scientific explanations: Focusing on the interplay among three levels of pictorial representation. *Int. J. Sci. Educ.* **2020**, *42*, 1–26. [CrossRef]
51. Stern, R.J. When and how did plate tectonics begin? Theoretical and empirical considerations. *Chin. Sci. Bull* **2007**, *52*, 578–591. [CrossRef]
52. Aufschnaiter, C.; Rogge, C. How Research on Students' processes of concept formation can inform curriculum development. In *The World of Science Education: Science Education Research and Practice in Europe*; Sense Publishers: Rotterdam, The Netherlands, 2012; pp. 63–90.
53. Bergold, J.; Thomas, S. Participatory research methods: A methodological approach in motion. *Hist. Soc. Res. /Hist. Soz.* **2012**, *37*, 191–222.
54. Kolb, S.M. Grounded theory and the constant comparative method: Valid research strategies for educators. *J. Emerg. Trends Educ. Res. Policy Stud.* **2012**, *3*, 83–86.
55. Tang, K.-S.; Won, M.; Treagust, D. Analytical framework for student-generated drawings. *Int. J. Sci. Educ.* **2019**, *41*, 2296–2322. [CrossRef]
56. Jindal-Snape, D.; Topping, K. Observational analysis within case study design. In *Using Analytical Frameworks for Classroom Research*; Rodrigues, S., Ed.; Routledge: London, UK, 2010; Volume 1, pp. 19–37.

Article

A Novel Modelling Process in Chemistry: Merging Biological and Mathematical Perspectives to Develop Modelling Competences

Vanessa Lang ^{1,*}, Christine Eckert ², Franziska Perels ², Christopher W. M. Kay ^{1,3} and Johann Seibert ^{1,*}

¹ Physical Chemistry and Didactics of Chemistry, University of Saarland, Campus B 2.2, 66123 Saarbrücken, Germany; christopher.kay@uni-saarland.de

² Department of Educational Sciences, University of Saarland, Campus A 4.2, 66123 Saarbrücken, Germany; christine.eckert@uni-saarland.de (C.E.); f.perels@mx.uni-saarland.de (F.P.)

³ London Centre for Nanotechnology, University College London, London WC1H 0AH, UK

* Correspondence: vanessa.lang@uni-saarland.de (V.L.); johann.seibert@uni-saarland.de (J.S.)

Abstract: Models are essential in science and therefore in scientific literacy. Therefore, pupils need to attain competency in the appropriate use of models. This so-called model-methodical competence distinguishes between model competence (the conceptual part) and modelling competence (the procedural part), wherefrom a definition follows a general overview of the concept of models in this article. Based on this, modelling processes enable the promotion of the modelling competence. In this context, two established approaches mainly applied in other disciplines (biology and mathematics) and a survey among chemistry teachers and employees of chemistry education departments (N = 98) form the starting point for developing a chemistry modelling process. The article concludes with a description of the developed modelling process, which by its design, provides an opportunity to develop students' modelling competence.

Keywords: models; modelling competence; chemical education

Citation: Lang, V.; Eckert, C.; Perels, F.; Kay, C.W.M.; Seibert, J. A Novel Modelling Process in Chemistry: Merging Biological and Mathematical Perspectives to Develop Modelling Competences. *Educ. Sci.* **2021**, *11*, 611.

<https://doi.org/10.3390/educsci11100611>

Academic Editors: Moritz Krell,
Andreas Vorholzer and
Andreas Nehring

Received: 9 August 2021

Accepted: 29 September 2021

Published: 3 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In societies that rank science and technology as highly important, enhancing pupils' participation is becoming increasingly central to teaching and learning strategies. Scientific reasoning represents an essential part of modern society since it incorporates contemporary philosophical and empirical psychological perspectives of science [1] and thereby enhances personal, social, professional and cultural participation [2]. The following six perspectives characterize the scientific reasoning: (1) postulation, (2) deployment of experiments both to control postulation and to explore observations, (3) hypothetical construction of analogical models, (4) structuring the natural variety by comparison and taxonomy, (5) statistical analysis of regularities of populations, and historical derivation of explanations [3]. Due to the central role of models in science practice and experimental studies, models are the primary method in science and scientific reasoning [2,4]. Besides appreciating the characteristics of models, their usage also belongs to scientific literacy as well [5]. Foremost in chemistry, models are essential tools for understanding and communication [6]. The great importance of models is mainly due to the nature of chemistry as a primarily abstract discipline [7]. This aspect arises from the fact that, in addition to the real visible macroscopic perspective, chemistry must consider sub-microscopic (atoms, molecules) and representational (equations, symbols) perspectives [8,9]. These characteristics lead to difficulties in the learning process for many pupils, especially when transitioning between different perspectives [10]. In contrast to the broad acceptance and importance of models and model competency in chemistry education, practice does not appear to address this issue adequately. For example, when using models in the classroom, the focus is often on describing them instead of predicting phenomena or solving problems. Furthermore,

student-centred modelling is seldom anchored in practice [11]. In addition, chemistry lessons often integrate the historically oriented development of particle models from undifferentiated over less differentiated (Shell model of atoms) to strongly differentiated models (orbital models). However, this can lead to learning difficulties [12] if pupils do not competently deal with the model concept. However, properly guided, this constant development offers potential for the acquisition of competencies in chemistry.

This article addresses this issue in several steps. We start with the theoretical foundation of models and pupils' model competences in chemistry education. Subsequently, a comparison of two modelling processes to promote the modelling competence of pupils is made. Finally, we present a blended approach for a modelling process in chemistry. The results of a survey among chemistry teachers (practice-oriented relationship to chemistry) and employees of chemistry education departments (research-oriented relationship to chemistry) support the argumentation.

2. Theoretical Foundations

2.1. The Concept of Models

Models are objects or theoretical constructs created or used by a subject for a specific purpose [13]. Certain properties of the model are associated with particular properties of the represented object [14]. Thus, models do not necessarily represent a complete picture of reality but often a specific aspect [15]. Thus, the modelling process of reality under different points of view, results in various types of models. The categorization arises, for example, from the function of the models (research model versus demonstration model, [16]) or the nature of the models (virtual vs. tangible, [17]). This article considers the following two categories in more detail as they appear in the intended chemical modelling process. Pedagogical analogical models share information with the represented object. Teachers or pupils create them to explain phenomena that are not accessible to people. One or more attributes usually dominate the structure of the model to underpin the explanation [18]. Learners generate mental models within their cognitive activity during modelling processes as mental representations to describe, explain or predict phenomena [19]. When working with models in science, they take on three different functions: models are used to describe, explain or predict chemical phenomena [20]. Models function as tools to acquire knowledge or forecasting tools. In addition, they can serve as learning aids [21] by breaking down anthropomorphic ideas, reducing complex connections, generalizing circumstances or illustrating chemical and mathematical-logical processes [13]. According to previous research, analogical models enhance scientific learning if used effectively [22]. Nevertheless, the appropriate use of models is a complex cognitive activity [23]. To master this activity successfully, pupils need to acquire so-called model competence [24].

2.2. Model Competence

Model competence used in biological contexts is the ability "to gain purposeful new insights into . . . topics using models, to judge models concerning their purpose, and to reflect on the epistemological process using models" [25] (p. 55). According to the second and third lines of Table 1, model competence has two sub-dimensions: knowledge about models and modelling. 'Knowledge about models' covers the conceptual part of competence and subsumes the 'nature of models' and 'multiple models'. In the dimension 'nature of models', learners compare the model and the represented object in terms of similarities and differences. In the context of 'multiple models', reasons for the existence of different models of a specific object are discussed. 'Modelling' as the procedural part of the model competence summarises 'purpose of models', 'testing models' and 'changing models'. The purposes merged under the category 'purpose of models' are general reasons for existing models and reasons to create and apply models (e.g., construction and evaluation of experiments, justification of causal relationships). Thus 'testing models' involves integrating different perspectives into the model, whereas a hypothesis may be tested by using the model. The formulated three levels of competence for each of these sub-

categories (Level I: exclusive consideration of the model; Level II: factual explanation of the phenomenon to generate understanding; Level III: hypothetical–deductive investigation of the phenomenon) allows a classification of the learners’ proficiency levels [26].

Table 1. Model–methodological competence [extension of 25].

Model-Methodological Competence				
Model Competence Knowledge about Models (Conceptual)		Modelling Competence Modelling (Procedural)		
nature of models	multiple models	purpose of models	testing models	changing models

Here, we propose a further differentiation between model competence and modelling competence to emphasize the procedural character to a greater extent. Modelling competence is the ability to initiate a theory-guided or creative process of cognition when creating models, to gain knowledge related to purpose when using models, make judgements about models regarding their purpose, and to reflect on the process of acquiring knowledge through models and modelling [27]. Comparable to the model competence (2nd and 3rd line of Table 1), the new definition specifies the epistemological procedures into a theory-based orientation and creative development. As a result, the sub-dimensions show four competence levels each [28]:

- Level I: Exclusive consideration of the model;
- Level II: Factual explanation of the phenomenon to generate understanding;
- Level IIIA: Abductive reasoning explanation of the phenomenon;
- Level IIIB: Hypothetical-deductive investigation of the phenomenon.

Following, but not in entire agreement with this definition, this paper defines model-methodical competence as an umbrella term of modelling competence (previously called ‘knowledge about models’, conceptual part) and the modelling competence (previously called ‘modelling’, procedural part) (light green terms in Table 1). The sub-dimensions remain in their allocation in the new definition. This distinction emerges from the differentiation between practical and meta-modelling knowledge [29].

2.3. Modelling Processes

The deep rootedness of models in science education described above emphasizes the promotion of modelling competence as a central task of chemistry education [30]. Thoughtful consideration of modelling processes as iterative cycles of creating, applying and reviewing models enables competence promotion [31]. In this context, students’ active manipulation of models positively affects model competence in three ways [32]: First, hands-on experience with models enables a cognitive off-load. Three-dimensional representations can spare cognitive capacities. Second, the pupils perceive multiple representations as they revise previous representations themselves or see the representations of their fellow pupils. The integration of various representations allows the learners to create more comprehensive and coherent mental models. Third, the physical confrontation with the model encounters various stimuli (such as cues from the touch) in the long-term memory, forming more bonds between the learning content and the long-term memory. This approach becomes even more critical against the background that pupils have not developed model competence in the sense mentioned above in a satisfactory way [33].

2.4. Modelling Process 1: Formally Used in Biological Contexts

The first modelling process formally used in biological contexts (Figure 1) starts with an experiment or a daily observation [34]. The data obtained on this basis influence the following preliminary considerations. Based on this, a model is generated in creative development, used to create a hypothesis. The hypothesis is then either verified or falsified

by further data. In the last case, the cycle is rerun using new data until the hypothesis can finally be accepted.

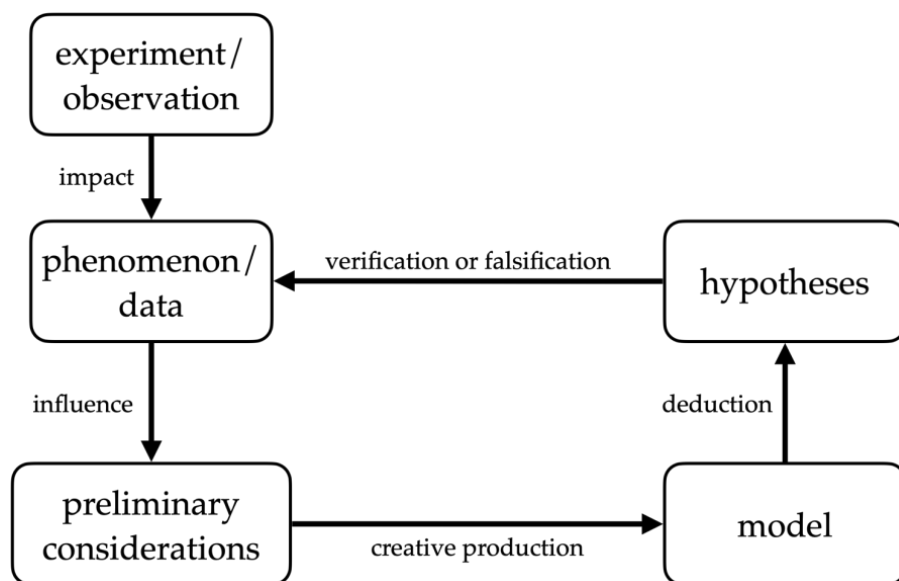


Figure 1. Scheme to promote model competence [34].

2.5. Modelling Process 2: Formally Used in Mathematical Contexts

The modelling circle shown in Figure 2 is formally used in mathematical contexts and reduced to the essentials. This process distinguishes in a 2×2 design between the world and mathematics as well as between the resulting problem and its mathematical solution [35]. In the chemical context of modelling, mathematics is equal to the model world. The starting point of this modelling process is an outer-mathematical problem (situation). A mathematical model with an inner-mathematical problem is generated from the situation by (mathematical) modelling. Applying mathematical rules and procedures causes an inner-mathematical consequence from the problem. In the next step, the modeller relates the mathematical results to the real world to obtain and check the plausibility of the mathematical results. If the result is not considered a valid answer for the initial situation, the process is rerun.

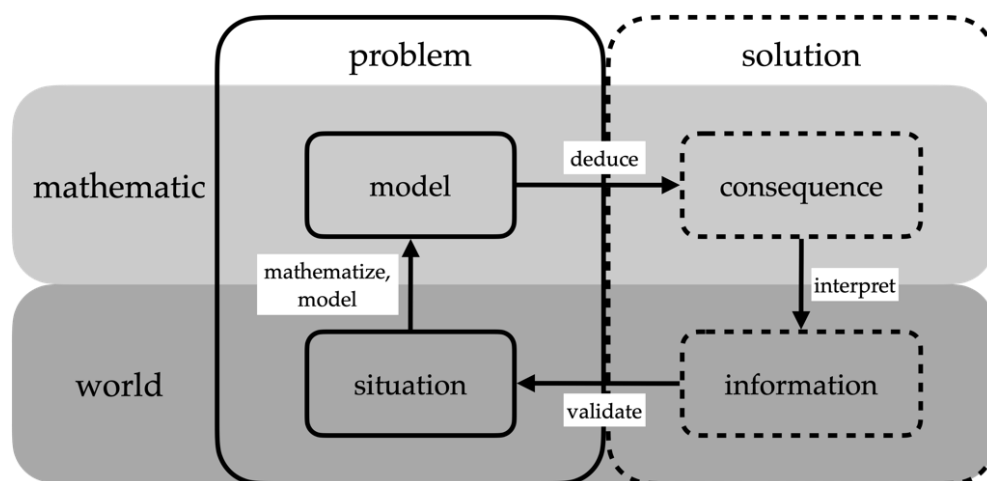


Figure 2. Mathematical modelling process [35].

3. Method

3.1. Research Questions

Based on the presented relevance of models and model competence for chemistry teaching, the online questionnaire covers the experiences with models (R 1) and model competence (R 2) of didactical experts in chemistry (chemistry teachers with a practice-oriented relationship to chemistry and employees of chemistry education departments with a research-oriented relationship to chemistry). Table 2 provides an overview of the research questions.

Table 2. Overview of the research questions.

Models in Teacher Training, Chemical Education Research or Chemistry Teaching
R 1. Which aspects related to models do anchor in teacher training, chemical education research or chemistry teaching?
Model competence
R 2a. To what extent is a well-developed model competence perceived as important for chemistry education?
R 2b. To what extent do the respondents agree with the dimensions of model competence [19] (Table 1) in the chemical context?
Modelling processes
R 3a. How do the respondents assess the transferability of modelling processes from other disciplines to chemistry in general?
R 3b. To what extent can the presented modelling process 1 (Figure 1) be transferred to chemistry?
R 3c. To what extent can the presented modelling process 2 (Figure 2) be transferred to chemistry?
R 4. Considering theoretical aspects and expert opinions, how can a novel modelling process for chemistry look like?

Modelling processes provide a method for promoting model competence [24]. These processes vary in their focus depending on the subject area (modelling processes 1 and 2 in Sections 2.4 and 2.5). Based on this, the respondents estimate to what extent approaches from other disciplines are transferable to modelling in chemistry (R 3a). Subsequently, the questionnaire presents the two modelling processes from Sections 2.4 and 2.5, and the participants precisely assess their suitability for chemistry (R 3b + c). The survey ultimately aims to design a novel chemistry modelling process that considers theoretical aspects and expert opinions (R 4).

3.2. Questionnaire

The first questions of the questionnaire collect background data. In addition to the relationship to chemistry (practice-oriented or research-oriented), this includes gender, age, teaching qualification (primary school, community school, vocational school, secondary school or other), location of study, school subjects, professional experience and research areas (for employees of chemistry education departments). Conforming to the research questions, the central part of the questionnaire consists of the sections: Models, Model competence and Modelling processes. Table 3 compiles an overview of the questions and associated answer options in the central part of the questionnaire. The questions in the Models section (Q1 to Q3, Table 3) capture the degree to which education, chemistry education research or chemistry education integrate models and which aspects they explicitly address. The subsequent part on model competence first presents the definition of model competence (Table 1). The respondents then evaluate this definition regarding its transferability to chemistry and give their subjective assessment of the importance of a strong model competence for chemistry (Q4, Table 3). Finally, the respondents name aspects of

the definition that fit well for chemistry and what should be omitted, added or changed if necessary (Q5 and Q6, Table 3). The last part about modelling processes starts with a general assessment of whether the respondents could imagine transferring modelling processes from biology or mathematics to chemistry (Q7 a and b, Table 3). After a video-based presentation of the first modelling process, the participants specifically indicate how to transfer to chemistry. In three open-ended questions, respondents then identify aspects that transfer well to chemistry, modify or omit, or add to better fit chemistry modelling (Q8 and Q9, Table 3). The final questions regarding the mathematical modelling process (Figure 2) are the same as those concerning the biological modelling process before (Q10 and Q11, Table 3).

Table 3. Overview of the central part of the questionnaire.

Models
<p>Q1. Did you get to know different models in chemistry (cf. particle models, molecule kits, model experiments) during your education?</p> <ol style="list-style-type: none"> 1. "No, I have never got to know different models in chemistry." 2. "Yes, during university education." 3. "Yes, during the second phase of training (preparatory service)." 4. "Yes, during in-service training or seminars." 5. "Yes, in the course of the following measure:" (open-ended answer)
<p>Q2. Do you/Does your research group explore different models in your/their research (cf. particle models, molecule kits, model experiments)? (research-orientated participants only)</p> <ol style="list-style-type: none"> 1. "No, I do not explore/ my research group does not explore different models." 2. "Yes, I explore different/ my research group explores models theoretically." 3. "Yes, I explore different/ my research group explores models practically concerning the use of different models, cf. by developing and testing possible applications for everyday school life." 4. "Yes, I explore different/ my research group explores different models in an inferential way:" (open-ended answer)
<p>Q3. In your chemistry lessons, do you employ different models (cf. particle models, molecule kits, model experiments)? (practice-orientated participants only)</p> <ol style="list-style-type: none"> 1. "No, I do not employ models in my lessons." 2. "Yes, I employ theoretical models (cf. particle models) in my lessons." 3. "Yes, I employ analogue models (cf. molecule kits) in my lessons." 4. "Yes, I employ conceptual models in my lessons (cf. mental representations of chemical laws such as the law of conservation of mass) to make and check predictions." 5. "Yes, I employ model experiments in my lessons (cf. Stechheber experiment)." 6. "Yes, I employ models in my lessons as follows:" (open-ended answer)
Model competence
<p>Definition of Model competence (Table 1)</p>
<p>Q4. Please indicate how much you think the following statements are true.</p> <ol style="list-style-type: none"> (a) "For me, this definition of model competence applies just as well to chemistry." 4-point Likert scale ranging from 1 'disagree' to 4 'agree'. (b) "For me, I see a well-developed model competence of the students according to the definition mentioned above as very important for chemistry." 4-point Likert scale ranging from 1 'disagree' to 4 'agree'.
<p>Q5. Which of the above aspects would you apply to model competence in chemistry? open-ended answer</p>
<p>Q6. Which aspects would you modify, add or omit for better applicability to model competence in chemistry? open-ended answer</p>
Modelling processes
<p>Q7. Please indicate how much you think the following statements are true.</p> <ol style="list-style-type: none"> (a) "In my opinion, modelling processes from biology are good transferrable to chemistry." 4-point Likert scale ranging from 1 'disagree' to 4 'agree'. (b) "In my opinion, modelling processes from mathematics are good transfer-able to chemistry." 4-point Likert scale ranging from 1 'disagree' to 4 'agree'.

Table 3. Cont.

Models
Video-based presentation of Modelling process 1
Q8. Please indicate how much you think the following statements are true. “In my opinion, modelling processes from biology are good transferrable to chemistry.” 4-point Likert scale ranging from 1 ‘disagree’ to 4 ‘agree.’
Q9. Which aspects of the modelling process from biology ... (a) “... fit well and are adaptable for chemistry, in your opinion?” <i>open-ended answer</i> (b) “... do not fit well and should be modified or omitted for chemistry, in your opinion?” <i>open-ended answer</i> (c) “... in your opinion would have to be supplemented for the scheme to represent a modelling process in chemistry?” <i>open-ended answer</i>
Video-based presentation of Modelling process 2
Q10. Please indicate how much you think the following statements are true. “In my opinion, modelling processes from mathematics are good transferrable to chemistry.” 4-point Likert scale ranging from 1 ‘disagree’ to 4 ‘agree.’
Q11. Which aspects of the modelling process from mathematics ... (a) “... fit well and are adaptable for chemistry, in your opinion?” <i>open-ended answer</i> (b) “... do not fit well and should be modified or omitted for chemistry, in your opinion?” <i>open-ended answer</i> (c) “... in your opinion would have to be supplemented for the scheme to represent a modelling process in chemistry?” <i>open-ended answer</i>

3.3. Participants

A total of 98 participants completed the qualitative questionnaires via Unipark. Among them, 56 (57%) were completed by research-orientated people (University) and 42 (43%) by practice-orientated people (School), with a total of 44 (44,9%) female, and 52 (53,1%) male (2 abstentions). The mean age amounts to 40.18 years (SD = 14.76), and 78 out of 98 (80%) respondents hold secondary school teaching qualifications. The most common teaching subject besides chemistry is biology (37 out of 93, 38%). Among the research-oriented participants, the number of people without professional experience of teaching at school is high (25 out of 56, 45%). In comparison, this proportion is only 17% (7 out of 42) among the practice-oriented participants. Table 4 shows an overview of the background data.

Table 4. Background data of the survey.

	In Total (n = 98)	Research-Orientated Relationship (n = 56)	Practice-Orientated Relationship (n = 42)
Gender	m = 52; f = 44	m = 31; f = 23	m = 21; f = 21
Age	M = 40.18; SD = 14.8	M = 39.02; SD = 15.5	M = 41.75; SD = 13.7
Teaching qualification	ps = 4; cs = 6; vs. = 2; ss = 78; o = 8	ps = 2; cs = 4; vs. = 2; ss = 43; o = 5	ps = 2; cs = 2; vs. = 0; ss = 35; o = 3
Location of study	bw = 9; by = 11; b = 5; bb = 1; hh = 1; he = 5; mv = 2; n = 5; nrw = 20; rlp = 5; sl = 18; s = 1; sa = 1; sh = 4; t = 3; o = 3	bw = 3; by = 9; b = 4; he = 4; mv = 1; n = 4; nrw = 15; rlp = 3; sl = 1; s = 1; sa = 1; sh = 3; t = 3; o = 2	bw = 6; by = 2; b = 1; bb = 1; hh = 1; he = 1; mv = 1; n = 1; nrw = 5; rlp = 2; sl = 17; sh = 1; o = 1
School subjects (selection)	biology = 35 mathematics = 18 physics = 11 Science = 8	biology = 19 mathematics = 8 physics = 8 Science = 3	biology = 16 mathematics = 10 physics = 3 Science = 5

Table 4. Cont.

	In Total (<i>n</i> = 98)	Research-Orientated Relationship (<i>n</i> = 56)	Practice-Orientated Relationship (<i>n</i> = 42)
Professional experience at school	M = 11.27; SD = 13.2 (0 years: 32 of 98)	M = 9.98; SD = 14.7 (0 years: 25 of 56)	M = 12.85; SD = 11.2 (0 years: 7 of 42)
Research Area (selection) <i>n</i> = 56	X	Digitalization = 20 Teacher education = 16 Experiments = 15 Models = 10	X

Key: m, male; f, female; ps, primary school; cs, community school; vs, vocational school; ss, secondary school; o, others; bw, Baden-Württemberg; by, Bayern; b, Berlin; bb, Brandenburg; hb, Bremen; hh, Hamburg; he, Hessen; mv, Mecklenburg-Vorpommern; n, Niedersachsen; nrw, Nordrhein-Westfalen; rlp, Rheinland-Pfalz; sl, Saarland; s, Sachsen; sa, Sachsen-Anhalt; sh, Schleswig-Holstein; t, Thüringen.

4. Results

4.1. Models

In the first section of the survey, the participants state that they all had experience of various models and mostly during their university education (72 out of 98; A in Figure 3). The chemistry education research models are rarely present, while theoretical and practical research approaches are approximately equally widespread (26 and 24 out of 56; B in Figure 3). In chemistry lessons, all the participants reported that they implement models in their teaching. Theoretical models occur to the same extent as analogue models (40 out of 42, 95.3%). Nevertheless, the teachers also use conceptual models (36 out of 42) and model experiments (34 out of 42, C in Figure 3).

4.2. Model Competence

Concerning model competence, the respondents not only considered model competence to be important for chemistry (MD = 3.40 of 4), but also could imagine transferring the model competence dimensions to their work (MD = 3.37 of 4; D & E in Figure 3). Most participants agreed that all the aspects to apply to chemistry (58 out of 98, F in Figure 3). The 'nature of models' was seen as critical for chemistry by 14 out of 98 people, with six participants with a research background stating the reason for this was that the initial object in chemistry is not empirical compared to objects in biology. These statements are consistent with the fact that in chemistry, frequently used models are models themselves, objects may be directly employed [10]. For example, chemists apply atomic models to explain macroscopic phenomena, mainly because the atomic structure is not observable directly. Therefore, modelling in chemistry operates on a different level. Furthermore, eight out of 98 respondents criticized the sub-dimension 'multiple models' with the explanation of a low significance of this dimension in the chemistry classroom (Q6, Table 5). Certain participants suggested adding types of models to the sub-dimension 'nature of models' or renaming 'changing models' to 'expanding models'. These results indicate that this competence definition is suitable for chemistry, but also that certain refinements are necessary.

4.3. Modelling Processes

In the previous section on modelling processes, the respondents rated the transferability of the modelling processes as neutral (2.85 (biology) and 2.58 (mathematics) on a 4-point Likert scale ranging from 1 'disagree' to 4 'agree'; G & H, Figure 3). After a video-based introduction of the modelling process formally used in biological contexts, the respondents rate the transferability of this process as rather good (M = 3.27 of 4; SD = 1.1; I in Figure 3). The participants indicated in open formats which aspects of the process fit well, which work more poorly and what the participants would probably like to add (Q9 a, Table 5). 61% (47 out of 77) agreed with all of the aspects, while 14 of 77 (18%) positively highlighted the experiment and 11 of 77 (14%) the formation of a hypothesis.

28 out of 65 (43%; Q9 b, Table 5) could not identify any aspect that should be changed or omitted in their opinion, though 6 out of 65 (9%) were critical of ‘creative production’. One criticism was, for example, that other aspects (cf. available resources) influence the ‘creative production’ besides the preliminary considerations. Concerning supplements, 14 out of 60 (23%) did not indicate anything (Q9 c, Table 5), while 10 out of 60 (17%) would like to add different levels of representation. This requirement matches the literature [8,36]. The second modelling process formally used in mathematical contexts takes this into account.

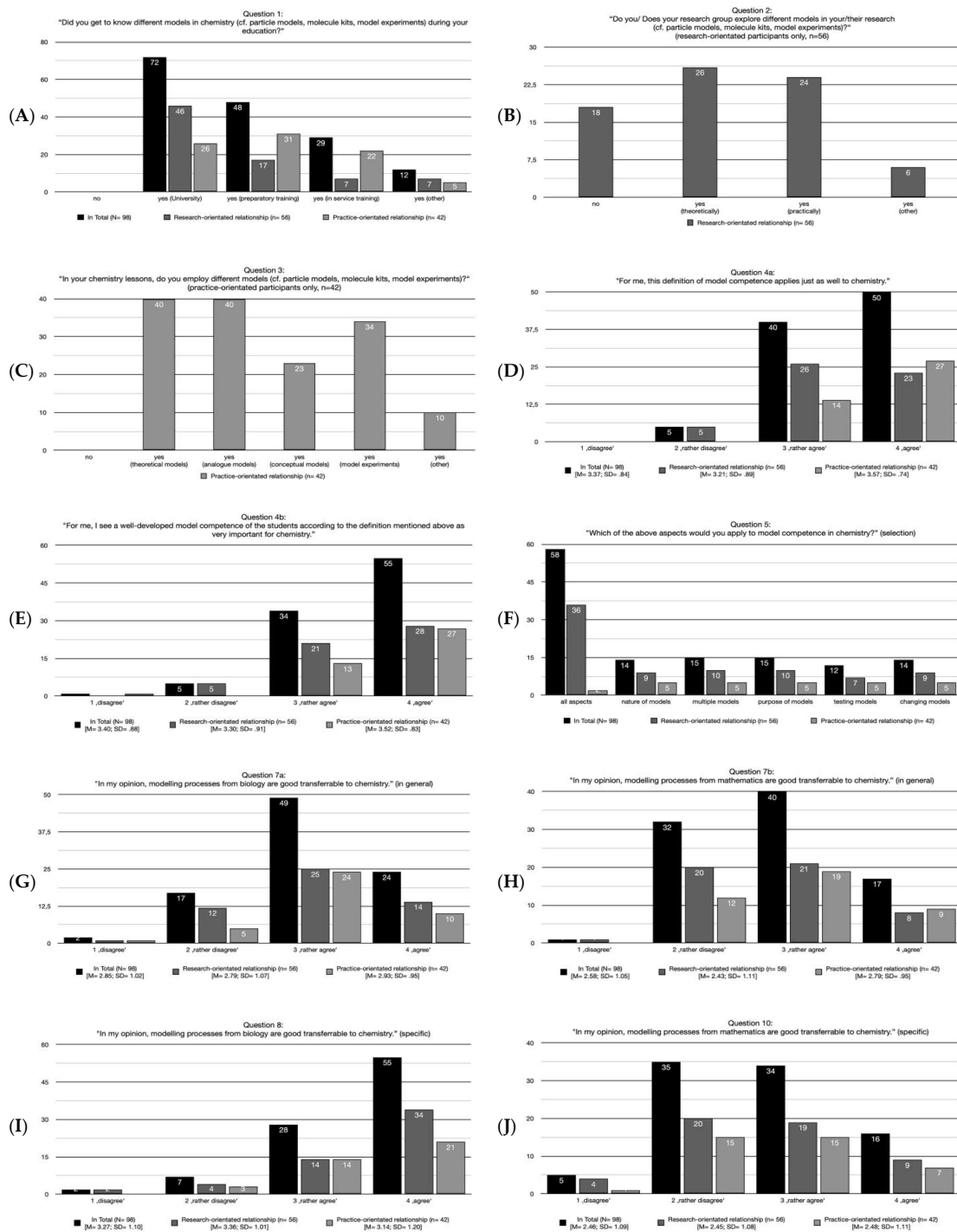


Figure 3. Results of the closed-ended questions. Diagrams showing the results of (A) question 1, (B) question 2, (C) question 3, (D) question 4a, (E) question 4b, (F) question 5, (G) question 7a, (H) question 7b, (I) question 8 and (J) question 10.

Table 5. Results of the open-ended questions.

Questions According to Table 3	In Total (N = 98)	Research-Orientated Relationship (n = 56)	Practice-Orientated Relationship (n = 42)
Model competence			
Q6 (selection)	n = 14; m = 8; p = 3; t = 7; c = 7	n = 11; m = 5; p = 3; t = 6; c = 3 "initial object unknown in chemistry" = 6 "Add types of models" = 2	n = 3; m = 3; p = 0; t = 1; c = 4 "rename 'changing models' to 'expanding' models" = 2
Modelling Processes			
Q9 (selection)	(a) a = 47; e = 14; h = 11; c = 7 (b) no = 28; pc = 6; cd = 6; m = 3; e = 4 (c) no = 14; lr = 9; e = 7; mm = 3	(a) a = 33; e = 6; h = 6; c = 3 (b) no = 17; pc = 4; cd = 4; m = 3 (c) no = 10; lr = 5; e = 5; mm = 3	(a) a = 14; e = 8; h = 5; c = 4 (b) no = 11; pc = 2; cd = 2; e = 4 (c) no = 4; lr = 5; e = 2
Q11 (selection)	(a) a = 8; dwm = 12; v = 10; m = 6; co = 6 (b) a = 2; no = 2; ma = 11; clr = 4 (c) no = 4; amp1 = 22; bm = 2	(a) a = 6; dwm = 10; v = 7; m = 6 (b) a = 2; no = 2; ma = 7 (c) no = 2; amp1 = 14; bm = 2	(a) a = 2; dwm = 2; v = 3; co = 6 (b) ma = 4; lr = 4 (c) no = 2; amp1 = 8

Key: Q9: a, all; e, experiment; h, hypotheses; c, cycle; no, nothing; pc, preliminary considerations; cd, creative development; m, models; lr, levels of representation; mm, mathematical models. Q11: a, all; dwm, distinction between the world and mathematics; v, validation; m, modelling; co, consequence; no, nothing; ma, mathematization; de, deduction; i, interpretation; lr, levels of representation; amp1: aspects of modelling process 1; bm, blended model.

The structure of this part of the questionnaire is analogous to the first modelling process: at the beginning, a video represents the process to the participants. Afterwards, they rate the transferability of this process as neutral ($M = 2.46$ of 4; $SD = 1.1$; J in Figure 3). The respondents were asked subsequently in an open form to indicate aspects that fit well, fit poorly and elements that probably need to be supplemented (Q11 a to c, Table 5). Eight out of 70 (11%) respondents saw all aspects as transferable to chemistry, while 17% (12 out of 70) perceived the juxtaposition of the 'world' and 'mathematics' (model world) as highly effective. Another 10 out of 70 (14%) highlighted referring to the real situation ('validate') as positive. The validating step aligns with the general orientation of chemical education towards everyday phenomena (cf. the processing circuit of chemistry in context, [37]). Significantly few participants (2 out of 56, 4%) stated that they would not change the presented aspects. Eleven of 56 (20%) named 'mathematize' as needing change, as mathematization rarely matters in the chemical modelling process. Among the supplementary proposals, 20 out of 55 participants (37%, Q11 in Table 5) mention aspects of the biological approach. These are copying effects created by the order of the tasks [38]. Apart from that, the participants named several different aspects. However, these only occurred twice each: The process makes sense if there is a transition between macroscopic and sub-microscopic levels, which is not the case in every chemical modelling. There should be an experiment or theoretical basement integrated as the starting point, which takes the importance of the experiment as the second central method of science into account [2].

4.4. Suggestion for a Novel Modelling Process in Chemistry

- The following consequences for modelling in chemistry emerge from the previous explanations:
- The aspects of the first modelling process (Figure 1) generally remain unchanged;
- The second model (Figure 2) emerges that the process needs to differentiate between the macroscopic real world and the sub-microscopic modelled world;
- For clarifying the cognitive processes (cf. mental analogue models), a separation occurs between perceptual and non-perceptual modelling steps;

- The modelling process should integrate phases to improve model-methodical competencies at appropriate points.

The developed process for modelling in chemistry shown in Figure 4 distinguishes between two different levels: the real macroscopic and the sub-microscopic modelled level. This explicit separation enables the pupils to consciously switch between the macro and sub-micro worlds with their unique peculiarities and regularities. The modelling process starts with a phenomenon that pupils can observe in their everyday lives. This starting point considers the general didactic demand for relevance in chemistry lessons [39]. Conversely, it enables students to understand that they should develop a model [40]. The experiment or phenomenon provides (experimental) data or observations that depend strongly on personal factors (e.g., disciplinary knowledge, theories, attention, [41]). In the following step, the modeller activates their prior knowledge and conceptual model competence (sub-categories ‘nature of models’ and ‘multiple models’, cf. *n* and *M* in Figure 4) to form a mental model. Here, a transition takes place in two ways. There is a change from the experiential real world to the model world (in Figure 4: light grey or dark grey background). Conversely, visible (Figure 4: solid outlines) processes become invisible (Figure 4: dashed contours). Using inner (modelling competence, creativity) and outer resources (learning situation, available materials), the modeller generates an analogue representation out of the mental model, a so-called pedagogical analogical model [19]. Mental models are simply a stopover in forming an analogical model. Nonetheless, the discrepancy is essential in analysing pupils’ concepts because mental models and their analogous representations do not necessarily coincide. The analogical model subsequently allows hypothesis generation. In this step, the model world refers to reality, i.e., the learner must once again make a ‘world transition’. This step gives pupils an understanding of the competence dimension ‘purpose of models’. Within the macroscopic world, the pupils generate experimental settings or everyday phenomena that provide observations or data to verify or falsify the hypothesis. By testing the hypothesis, learners create a reference to the modelling competence dimension ‘testing models’. In the case of a falsified hypothesis, the modellers go through the cycle again. First, they change their mental model and thus also the analogical model. Through the model modification, the modelling process establishes a relationship to the competence ‘changing models’.

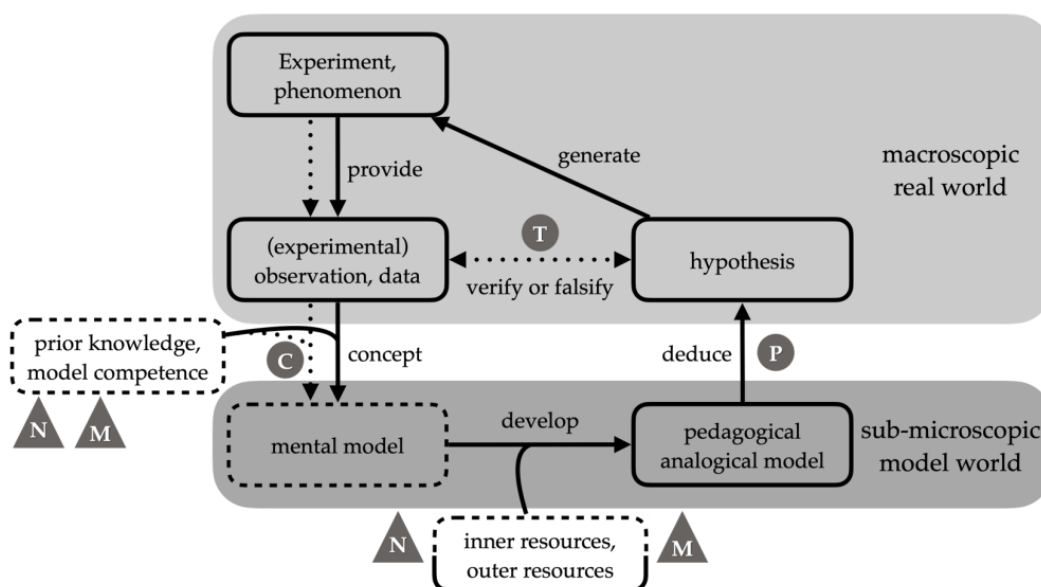


Figure 4. Modelling process in chemistry to promote modelling competence. Notice: The icons mark points to address the sub-competencies during the process (round icons)/applied (triangular icons) (N, nature, M, multiple, P, purpose, T, testing, C, changing).

A hypothetical example from school practice on states of matter serves to concretize the proposed modelling process for chemistry (Figure 5). The pupils observe the evaporation of water in their everyday life, for example, when cooking. From this, they derive data, the boiling temperature of the water or the optical properties of water vapour, for example. To explain these data, the pupils activate their prior knowledge of the undifferentiated particle model and their model competence. This activation generates a pictorial representation in the students' minds—a mental model (pictorial mental representation in Figure 5). Using coloured cardboard or other craft materials, the pupils visualize their mental representation to form a pedagogical analogue model (visualization in Figure 5). For example, circles cut out from cardboard could be arranged at small distances to represent the liquid state and at large distances for the gaseous state. In this step, personal skills (creativity, manual skills) play an essential role. The analogue model is now observable for the teacher and the fellow pupils. From the analogue model, the students then derive, for example, the hypothesis that a certain amount of substance would have to occupy a larger volume after the transition from the liquid to the gaseous state of matter since the movement of the particles increases and the particles occupy the entire available volume. To test the hypothesis, the students then perform the experiment on the evaporation of acetone to exemplify the transitions between the states of matter in a closed system (pointed arrows in Figure 5). They place a few millilitres of acetone into a plastic bag and close it airlessly. Using a hot water bath, the students then heat the acetone and see that the bag inflates. The hypothesis can therefore be accepted, which strengthens the students' model conception.

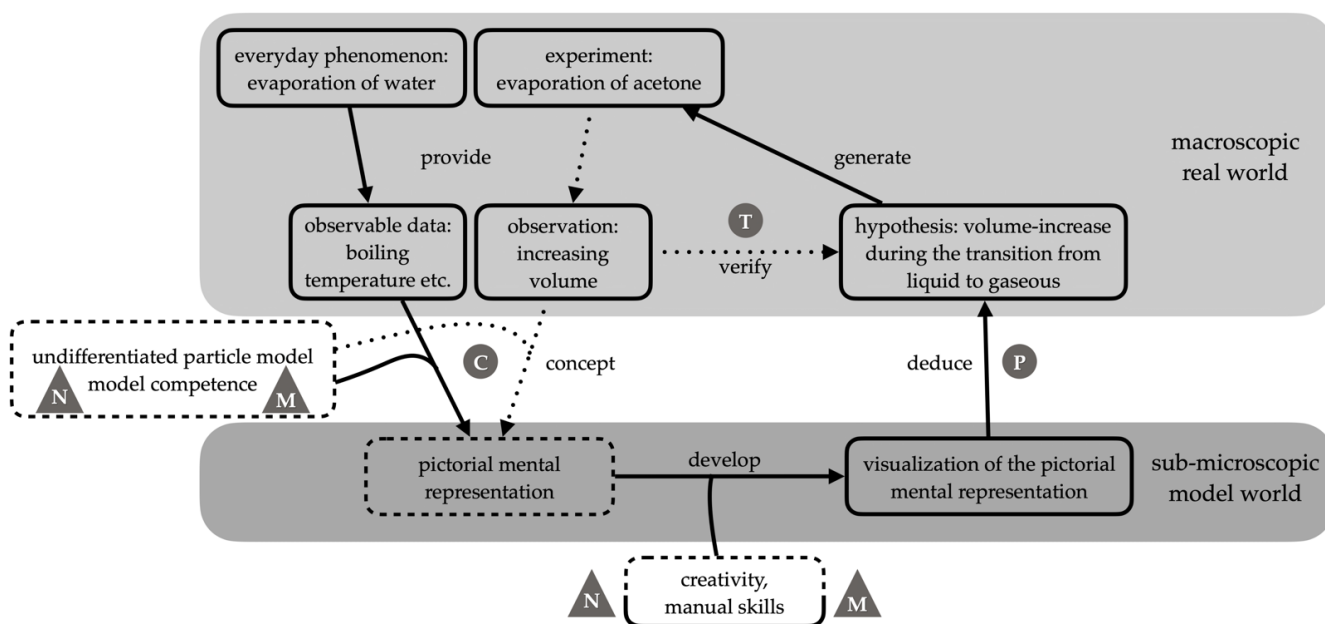


Figure 5. Hypothetical example of the modelling process in chemistry from school practice. Notice: The icons mark points to address the sub-competencies during the process (round icons)/applied (triangular icons) (N, nature, M, multiple, P, purpose, T, testing, C, changing).

5. Discussion

The following section discusses the developed model on the basis of accepted didactic concepts and classifies the idea correspondingly. In the SDDS approach, Klahr & Dunbar [42] anchor modelling in scientific reasoning. The approaches of Clement [43] and Göhner & Krell [44] support the basic structure. The basic structure consisting of conjecture, evaluation and modification or reflection of Clement [43] and Göhner and Krell [44] is common to the developed scheme. In addition, loops are integrated in all approaches, enabling a reference back to the question, hypothesis or built model. The developed scheme of a modelling process agrees with Göhner and Krell [44] in the dis-

inction between the macroscopic real world and the sub-microscopic model world. In this context, Steinbuch [45] describes that with every transition between the real and the model world, filtering always occurs, in which the subject only processes aspects that are considered important. Moreover, the developed model additionally distinguishes between observable and non-observable stages according to Harrison and Treagust's [18]. Fratiwi et al. additionally emphasise the importance of knowing students' mental models in order to assess their scientific understanding [46]. Therefore, the scheme includes mental models as well as analogue models. Mental models and formed analogue models often differ and consequently, the captured modelling competence can differ from the actual modelling competence. Moreover, Didiş, Eryılmaz and Erkoç [47] describe the basis for the formation of mental models as a combination of scientific and non-scientific fragments, whereby the developed schema incorporates prior knowledge and model competence. When creating the analogue model, the influencing factors (model competence, prior knowledge, external and internal resources) are again considered. Above this, the focus of the second part of the scheme (formation of hypotheses and seeking of verification or falsification) bases on the separation between search hypothesis space, test hypothesis and evidence evaluation of the SDDS approach [42]. At last, the novel scheme, based on the reflection scheme of Caspari et al. [48], establishes relationships to the competence dimensions at the appropriate points to force the promotion of modelling competence.

6. Conclusions

The blended process presented here brings the positive aspects of both modelling cycles together and compares them with the expert opinions from the survey. The design of an intervention will rest on this such that it is suitable for promoting modelling competence. The process will consider the complexity of modelling [23] and the pupils' attitudes by supporting them individually in their learning [49]. Unlike other studies that locate modelling processes in the upper secondary school [48,50], the intervention is anchored in initial chemistry teaching. Research has shown that misconceptions are stable over time and difficult to correct, including with increasing subject knowledge [51]. Nevertheless, the cognitive abilities of pupils increase during their school career, which means that the complexity of modelling processes can also rise to increase educational attainment [10]. Therefore, a spiral curricular promotion of modelling competence is apparent and suggested in the survey responses.

Author Contributions: Conceptualization, V.L. and J.S.; methodology, V.L. and C.E.; formal analysis, V.L.; investigation, V.L.; writing—original draft preparation, V.L. and J.S.; writing—review and editing, V.L., J.S. and C.W.M.K.; visualization, V.L.; supervision, C.W.M.K., J.S. and F.P.; project administration, V.L. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: The respondents agreed to data use for research.

Data Availability Statement: Data are available if required.

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

References

- Osborne, J. The 21st Century Challenge for Science Education: Assessing Scientific Reasoning. *Think. Ski. Creat.* **2013**, *10*, 265–279. [CrossRef]
- Kultusministerkonferenz Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss; Luchterhand: München, Neuwied, 2004.
- Kind, P.; Osborne, J. Styles of Scientific Reasoning: A Cultural Rationale for Science Education? *Sci. Ed.* **2017**, *101*, 8–31. [CrossRef]
- Giere, R.N. *Science without Laws*; Science and its conceptual foundations; University of Chicago Press: Chicago, IL, USA, 1999; ISBN 978-0-226-29208-3.
- Giere, R.N. Understanding and Evaluating Theoretical Hypotheses. In *Understanding Scientific Reasoning*; Thomson/Wadsworth: Belmont, CA, USA, 1991; pp. 12–39. ISBN 978-0-15-506326-6.

6. Emden, M.; Ropohl, M.; Sommer, K. Modellieren als Methode der Erkenntnisgewinnung Eine Prozess-Perspektive auf eine naturwissenschaftliche Arbeitsweise. *Unterr. Chem.* **2019**, *171*, 7–11.
7. Justi, R.; Gilbert, J. The role of analog models in the understanding of the nature of models in chemistry. In *Metaphor and Analogy in Science Education*; Aubusson, P., Harrison, A.G., Ritchie, S., Eds.; Science & Technology Education Library; Springer: Dordrecht, The Netherlands, 2006; pp. 119–130. ISBN 978-1-4020-3829-7.
8. Johnstone, A.H. The Development of Chemistry Teaching: A Changing Response to Changing Demand. *J. Chem. Educ.* **1993**, *70*, 701. [CrossRef]
9. Becker, H.-J.; Hildebrandt, H. Unanschauliches Veranschaulicht“-Modellexperimente Im Chemieunterricht Als Chance Für Analogiebildungen. *PdN-ChiS* **2003**, *52*, 15–19.
10. Coll, R.K. The role of models, mental models and analogies in chemistry teaching. In *Metaphor and Analogy in Science Education*; Aubusson, P., Harrison, A.G., Ritchie, S., Eds.; Springer: Dordrecht, The Netherlands, 2006; pp. 65–77. ISBN 978-1-4020-3829-7.
11. Nielsen, S.S.; Nielsen, J.A. Models and Modelling: Science Teachers’ Perceived Practice and Rationales in Lower Secondary School in the Context of a Revised Competence-Oriented Curriculum. *EURASIA J. Math. Sci. Tech. Ed.* **2021**, *17*, em1954. [CrossRef]
12. Eilks, I. Neue Wege zum Teilchenkonzept. Wie man Basiskonzepte forschungs- und praxisorientiert entwickeln kann. *Nat. Im. Unterricht. Chem.* **2007**, *18*, 23–27.
13. Gilbert, J.; Boulter, C.J. (Eds.) *Developing Models in Science Education*; Kluwer Academic Publishers: New York, NY, USA, 2000; ISBN 978-94-010-0876-1.
14. Mikelskis-Seifert, S.; Knittel, C.; Pfohl, U. Vom Modellieren Im Alltag Zum Modellieren Im Unterricht. *NiU* **2011**, *22*, 13–18.
15. Barke, H.-D.; Harsch, G.; Kröger, S.; Marohn, A. *Chemiedidaktik Kompakt*; Springer: Berlin/Heidelberg, Germany, 2018; ISBN 978-3-662-56491-2.
16. Schorn, J. Methoden und Modelle. In *Chemie-Methodik: Handbuch für die Sekundarstufe I und II*; Kranz, J., Schorn, J., Eds.; Cornelsen: Berlin, Heidelberg, Germany, 2012; pp. 162–173. ISBN 978-3-589-22379-4.
17. Nerdel, C. *Grundlagen der Naturwissenschaftsdidaktik*; Springer: Berlin/Heidelberg, Germany, 2017; ISBN 978-3-662-53157-0.
18. Harrison, A.G.; Treagust, D.F. A Typology of School Science Models. *Int. J. Sci. Educ.* **2000**, *22*, 1011–1026. [CrossRef]
19. Buckley, B.C.; Boulter, C.J. Investigating the Role of Representations and Expressed Models in Building Mental Models. In *Developing Models in Science Education*; Gilbert, J.K., Boulter, C.J., Eds.; Kluwer Academic Publishers: New York, NY, USA, 2000; pp. 119–135. ISBN 978-94-010-0876-1.
20. Franco, C.; Colinviaux, D. Grasping Mental Models. In *Developing Models in Science Education*; Gilbert, J.K., Boulter, C.J., Eds.; Kluwer Academic Publishers: New York, NY, USA, 2000; pp. 93–118. ISBN 978-94-010-0876-1.
21. Kircher, E. Zum Modellbegriff und zu seiner Bedeutung für den naturwissenschaftlichen Unterricht. In *Atommodelle im Naturwissenschaftlichen Unterricht*; Weninger, J., Brünger, H., Universität Kiel, Eds.; Beltz: Weinheim, Germany; Basel, Switzerland, 1976; pp. 248–263. ISBN 978-3-407-69112-5.
22. Harrison, A.G.; Treagust, D.F. Teaching and Learning with Analogies- Friend or Foe? In *Metaphor and Analogy in Science Education*; Aubusson, P., Harrison, A.G., Ritchie, S., Eds.; Science & Technology Education Library, Springer: Dordrecht, The Netherlands, 2006; pp. 11–24. ISBN 978-1-4020-3829-7.
23. Coll, R.K.; Lajum, D. Modeling and the Future of Science Learning. In *Models and Modeling*; Khine, M.S., Saleh, I.M., Eds.; Springer: Dordrecht, The Netherlands, 2011; pp. 3–21. ISBN 978-94-007-0448-0.
24. Schwarz, C.V.; White, B.Y. Metamodeling Knowledge: Developing Students’ Understanding of Scientific Modeling. *Cogn. Instr.* **2005**, *23*, 165–205. [CrossRef]
25. Grünkorn, J.; Upmeier zu Belzen, A.; Krüger, D. Design and Test of Open-Ended Tasks to Evaluate a Theoretical Structure of Model Competence. In *Authenticity in Biology Education-Benefits and Challenges*; Yarden, A., Ed.; University of Minho: Braga, Portugal, 2011.
26. Upmeier zu Belzen, A.; Krüger, D. Modellkompetenz Im Biologieunterricht. *ZfDN* **2010**, *16*, 41–57.
27. Krüger, D.; Upmeier zu Belzen, A. Kompetenzmodell der Modellierkompetenz–Die Rolle abduktiven Schließens beim Modellieren. *ZfDN* **2021**. [CrossRef]
28. Upmeier zu Belzen, A.; Engelschalt, P.; Krüger, D. Modeling as Scientific Reasoning—The Role of Abductive Reasoning for Modeling Competence. *Educ. Sci.* **2021**, *11*, 495. [CrossRef]
29. Schwarz, C.V.; Reiser, B.J.; Davis, E.A.; Kenyon, L.; Achér, A.; Fortus, D.; Shwartz, Y.; Hug, B.; Krajcik, J. Developing a Learning Progression for Scientific Modeling: Making Scientific Modeling Accessible and Meaningful for Learners. *J. Res. Sci. Teach.* **2009**, *46*, 632–654. [CrossRef]
30. Graf, E. Modelle Im Chemieunterricht. *NiU* **2002**, *13*, 4–9.
31. Koch, S.; Krell, M.; Krüger, D. Förderung von Modellkompetenz Durch Den Einsatz Einer Blackbox. *Erkenn. Biol.* **2015**, *14*, 93–108.
32. Stull, A.T.; Gainer, M.; Padalkar, S.; Hegarty, M. Promoting Representational Competence with Molecular Models in Organic Chemistry. *J. Chem. Educ.* **2016**, *93*, 994–1001. [CrossRef]
33. Lazenby, K.; Rupp, C.A.; Brandriet, A.; Mauger-Sonnek, K.; Becker, N.M. Undergraduate Chemistry Students’ Conceptualization of Models in General Chemistry. *J. Chem. Educ.* **2019**, *96*, 455–468. [CrossRef]
34. Upmeier zu Belzen, A.; Krüger, D. Ein Fall Für Erkenntnisgewinnung- Biologische Beiträge Zu Einem Verständnis Naturwissenschaftlichen Modellierens. *NiU* **2019**, *30*, 38–41.

35. Schupp, H. Anwendungsorientierter Mathematikunterricht in Der Sekundarstufe I Zwischen Tradition Und Neuen Impulsen. *Mathematikunterricht* **1988**, *34*, 5–16.
36. Gabel, D.; Briner, D.; Haines, D. Modelling with Magnets: A Unified Approach to Chemistry Problem Solving. *Sci. Teach.* **1992**, *59*, 58–63.
37. Nentwig, P.M.; Demuth, R.; Parchmann, I.; Ralle, B.; Gräsel, C. Chemie Im Kontext: Situating Learning in Relevant Contexts While Systematically Developing Basic Chemical Concepts. *J. Chem. Educ.* **2007**, *84*, 1439. [CrossRef]
38. Riese, J.; Reinhold, P. Entwicklung eines Leistungstests für fachdidaktisches Wissen. In *Methoden in der Naturwissenschaftsdidaktischen Forschung*; Krüger, D., Parchmann, I., Schecker, H., Eds.; Springer Spektrum: Berlin/Heidelberg, Germany, 2014; pp. 257–267. ISBN 978-3-642-37826-3.
39. Stuckey, M.; Sperling, J.P.; Hofstein, A.; Mamlok-Naaman, R.; Eilks, I. Ein Beitrag zum Verständnis der Relevanz des Chemieunterrichts. *CHEMKON* **2014**, *21*, 175–180. [CrossRef]
40. Grooms, J.; Fleming, K.; Berkowitz, A.R.; Caplan, B. Exploring Modeling as a Context to Support Content Integration for Chemistry and Earth Science. *J. Chem. Educ.* **2021**, *98*, 2167–2175. [CrossRef]
41. Eberbach, C.; Crowley, K. From Everyday to Scientific Observation: How Children Learn to Observe the Biologist's World. *Rev. Educ. Res.* **2009**, *79*, 39–68. [CrossRef]
42. Klahr, D.; Dunbar, K. Dual Space Search during Scientific Reasoning. *Cogn. Sci.* **1988**, *12*, 1–48. [CrossRef]
43. Clement, J. Learning via Model Construction and Criticism. In *Handbook of Creativity; Perspectives on Individual Differences*; Torrance, E.P., Glover, J.A., Ronning, R.R., Reynolds, C.R., Eds.; Plenum Press: New York, NY, USA, 1989; ISBN 978-0-306-43160-9.
44. Göhner, M.; Krell, M. Modellierungsprozesse von Lehramtsstudierenden Der Biologie. *Erkenn. Biol.* **2018**, *17*, 45–61.
45. Steinbuch, K. Denken in Modellen. In *Denken in Modellen*; Schäfer, G., Trommer, G., Wenk, K., Eds.; LEITTHEMEN Beiträge zur Didaktik der Naturwissenschaften; Westermann: Braunschweig, Germany, 1977; ISBN 978-3-14-167154-4.
46. Fratiwi, N.J.; Samsudin, A.; Ramalis, T.R.; Saregar, A.; Diani, R. Developing MeMoRI on Newton's Laws: For Identifying Students' Mental Models. *Eur. J. Educ. Res.* **2020**, *9*, 699–708. [CrossRef]
47. Didiş, N.; Eryılmaz, A.; Erkoç, Ş. Investigating Students' Mental Models about the Quantization of Light, Energy, and Angular Momentum. *Phys. Rev. ST Phys. Educ. Res.* **2014**, *10*, 020127. [CrossRef]
48. Caspari, I.; Weber-Peukert, G.; Graulich, N. Der Einsatz von Modellen Zum Erkenntnisgewinn-Eine Unterrichtseinheit Zur Förderung Der Modellkompetenz Im Kontext, Batterie“ Unter Explizitem Einbezug von Schülervorstellungen. *CHEMKON* **2018**, *25*, 23–34. [CrossRef]
49. Jansoon, N.; Coll, R.; Somsook, E. Understanding Mental Models of Dilution in Thai Students. *Int. J. Environ. Sci.* **2009**, *4*, 147–168.
50. Sarıtaş, D.; Özcan, H.; Adúriz-Bravo, A. Observation and Inference in Chemistry Teaching: A Model-Based Approach to the Integration of the Macro and Submicro Levels. *Sci. Educ.* **2021**, *30*, 1289–1314. [CrossRef]
51. Nicoll, G. A Report of Undergraduates' Bonding Misconceptions. *Int. J. Sci. Educ.* **2001**, *23*, 707–730. [CrossRef]

A Model of Scientific Data Reasoning

Amy M. Masnick ^{1,*} and Bradley J. Morris ²¹ Psychology Department, Hofstra University, Hempstead, NY 11549, USA² Educational Psychology Department, Kent State University, Kent, OH 44242, USA; bmorri20@kent.edu

* Correspondence: amy.m.masnick@hofstra.edu

Abstract: Data reasoning is an essential component of scientific reasoning, as a component of evidence evaluation. In this paper, we outline a model of scientific data reasoning that describes how data sensemaking underlies data reasoning. Data sensemaking, a relatively automatic process rooted in perceptual mechanisms that summarize large quantities of information in the environment, begins early in development, and is refined with experience, knowledge, and improved strategy use. Summarizing data highlights set properties such as central tendency and variability, and these properties are used to draw inferences from data. However, both data sensemaking and data reasoning are subject to cognitive biases or heuristics that can lead to flawed conclusions. The tools of scientific reasoning, including external representations, scientific hypothesis testing, and drawing probabilistic conclusions, can help reduce the likelihood of such flaws and help improve data reasoning. Although data sensemaking and data reasoning are not supplanted by scientific data reasoning, scientific reasoning skills can be leveraged to improve learning about science and reasoning with data.

Keywords: data reasoning; scientific reasoning; statistics education; numerical cognition; cognitive development; number sense

Citation: Masnick, A.M.; Morris, B.J. A Model of Scientific Data Reasoning. *Educ. Sci.* **2022**, *12*, 71. <https://doi.org/10.3390/educsci12020071>

Academic Editors: Moritz Krell, Andreas Vorholzer and Andreas Nehring

Received: 24 August 2021

Accepted: 12 January 2022

Published: 20 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Data reasoning is a critical skill in scientific reasoning. Although evidence evaluation is a step in many models of scientific reasoning (e.g., [1–3]), there has been much less attention on the interpretation of numerical data itself within this context, which has been investigated largely within the field of statistics [4]. We outline a model of data reasoning that describes how data sensemaking underlies data reasoning (both defined below). We further suggest that scientific data reasoning differs from both informal data reasoning and data sensemaking. We use the phrase scientific data reasoning to refer to a set of skills that help reasoners improve the quality of their data analysis and interpretation, which improves the quality of inferences that can be drawn from data. Although these skills are most commonly used in scientific reasoning contexts, they can be used in any context. Scientific data reasoning includes a set of skills that help to harness data sensemaking and strengthen everyday data reasoning by improving the systematicity of data collected via the scientific method, the quality of analysis via statistical tools, and inferences by reducing cognitive bias and providing tools for evaluating conclusions.

Science refers to both a body of knowledge and the processes that create and evaluate this knowledge [5]. These processes include generating and testing hypotheses, acquiring data, and evaluating theories with new data [6–8]. The cornerstone of the scientific process is the reliance on empirical data that are formally analyzed [5,8]. Research that has focused on understanding the cognitive processes that underlie scientific reasoning includes studies of generating hypotheses [1,5,9,10], making predictions [11], deciding how to measure variables [12–14], and interpreting data in light of theory and prior beliefs [2,3,15,16]. We focus on one area that has gotten less attention, specifically on how people make sense of numerical data. Golemund and Wickham [4] propose a cognitive interpretation of

data analysis, considering the sensemaking process of data, focused on how we reason with quantitative measurements. We argue that data reasoning is a unique skill, built on intuition and knowledge about math and statistics, and that a full understanding of scientific reasoning requires understanding of data reasoning as a core competency. In turn, this understanding has implications for teaching of these skills, and for future research directions.

We suggest that scientific reasoning is a specific type of information seeking that emerges from an acquired set of cultural tools and knowledge that augment general cognitive mechanisms (e.g., encoding, strategy use) and dispositions (e.g., curiosity [5,7,17,18]). Scientific reasoning is defined in multiple ways. Zimmerman and Klahr [5] describe scientific thinking as a multi-faceted, multi-level process that includes problem solving and reasoning, knowledge seeking, curiosity refined by science education that teaches hypothesis testing and evidence evaluation, and the development of metacognitive skills. Englemann et al. [19] describe scientific reasoning conceptualizations as falling into discussions of the process of scientific discovery, scientific argumentation, and understanding the nature of science. We position ourselves within Englemann et al.'s description of the process of scientific discovery, specifically focused on the evidence evaluation described by Zimmerman and Klahr, while still shaped by teaching, hypothesis testing, and general reasoning.

Data reasoning is situated within this definition in multiple places. One, informal data reasoning occurs frequently in daily life and requires no formal instruction (e.g., [19,20]). This is the process of drawing inferences from everyday data, requiring only general-purpose mechanisms for their operation (e.g., strategy acquisition and selection). Two, cultural tools relevant to data include number systems, mathematical formalisms, and data visualization tools (e.g., [21]), as well as the scientific process that produces (e.g., experimentation, data recording, as noted by many researchers, including [5,22]) and allows for analysis of data (e.g., inferential statistics, [4]). The acquisition of these formal tools improves the accuracy of data reasoning and improves the basic reasoning processes as well. Thus, we argue that understanding the cognitive processes underlying data reasoning is an important part of a fuller understanding of scientific reasoning.

We define data sensemaking as detecting trends or differences in sets of data without the use of formal analytical tools [4,23,24]. We draw evidence from perception [25] and numerical cognition [26] to suggest that data sensemaking relies on the same perceptual mechanisms that allow people to quickly aggregate and summarize complex information from the environment [27]. The perceptual processes and intuitions in data sensemaking work impressively well for detecting strong patterns in data without formal analysis (e.g., [3,28,29]). Detecting features in data provides information from which people can draw inferences and interpret the meaning of these features and patterns.

We define data reasoning as reasoning in which quantitative data are used as evidence for drawing conclusions or making decisions [30]. Data reasoning is a deliberate cognitive process that is limited by our memory [31] and our cognitive biases [4]. We argue that data sensemaking is a precursor to data reasoning, in that first the data are summarized (often via automatic perceptual processes), and then the summaries are used to draw inferences. Everyday data reasoning is common, though such inferences may fall prey to cognitive biases such as the tendency to confirm one's previous beliefs (i.e., confirmation bias, [32]).

Data reasoning occurs informally in many contexts. For example, when comparing the performance of players for a fantasy sports league, people do not need to conduct an experiment or test hypotheses. Instead, they are more likely to pay attention to features of the data, such as means or medians, variation, and trends, to guide their inferences or decisions (e.g., [20,30,33–36]). Researchers offer varied definitions of informal inferential reasoning (what we are calling data reasoning), but Makar and Rubin [30] highlight five key components that help describe the space: (1) informal reasoning makes a claim beyond the data; (2) conclusions are expressed with uncertainty; (3) data are used as evidence (i.e., the conclusions cannot be derived solely from prior beliefs, though the data can be considered

in theoretical contexts); (4) the data are considered in aggregate, and the reasoning stems from assessments of the aggregate data; and (5) context is considered.

In contrast, when engaging in scientific data reasoning, researchers use some of the same techniques as in data reasoning, but add safeguards to limit biases, which lead to the more deliberative scientific reasoning process. For example, scientists aim to ground research questions in theories backed by past evidence (e.g., [1,8,37]), design studies that measure and control variables to limit confounds (e.g., [38]), use external representations to represent a larger quantity of data at once (e.g., [21]), and apply formal statistical analyses to provide quantitative evidence when making inferences (e.g., [4]).

Our definitions of both scientific reasoning and scientific data reasoning are situated at a relatively coarse grain size. One piece of every model of scientific reasoning involves evidence evaluation, and that evidence evaluation typically includes evaluation and interpretation of quantitative data. Our model of data reasoning is at a somewhat more detailed level, including descriptions of component processes at a finer grain size than general models of scientific reasoning. One example of a description on a fine grain size is the process of summarizing data in sets. We propose that the same processes that underlie summarization in perceptual sets (e.g., dots) operate upon sets of numbers. These processes provide summaries that set up inferences from data. Thus, we provide descriptions at varying grain sizes across the scope of the review.

In this paper, we suggest a model of the cognitive processes underlying how people make sense of and draw inferences from data. We further suggest that, as people learn about science, they acquire tools to improve both of these processes. Our review targets cognitive processes as described from general Cognitive Science framework. We describe scientific reasoning as a process that includes declarative (e.g., scientific facts) and procedural (e.g., conducting unconfounded experiments) elements. We screened the literature to find research that was relevant to such a cognitive process model. We did not perform a literature search by term because the same term can have quite different meanings in different research traditions. Even within the community of scholars who study scientific reasoning, there is little consistency in the terminology that is used to describe it. For example, the special issue uses the term “competencies” [39] to describe what we and our colleagues often refer to as “processes” (e.g., [40]). Another reason for our selections is that we have attempted to bridge multiple literatures that have not regularly communicated, such as cognitive scientists, statisticians, and science educators. As Fischer et al. [6] note, “contemporary knowledge about what constitutes these competencies . . . is scattered over different research disciplines” (p. 29).

In sum, we propose that numerical data reasoning is rooted in intuitions about number sets and becomes more sophisticated as people acquire scientific and statistical reasoning skills. Although there are many types of nonquantitative data used in both everyday and scientific reasoning, in this paper, we focus exclusively on numerical data reasoning, and hereafter use the term data reasoning to refer to this type of reasoning. We argue that numerical data reasoning begins with data sensemaking, a largely intuitive process that summarizes sets of numbers much like summaries of perceptual features (e.g., relative size). Data sensemaking creates approximations of data rapidly. Data reasoning, or drawing conclusions from these data, is derived from these data summaries. Although these processes are fast and accurate given clear patterns or differences, both may be sensitive to cognitive limitations such as confirmation bias. Scientific data reasoning augments these informal processes by adding cultural tools for improving the accuracy of data gathering, representation, analysis, and inferences. We summarize our proposed model in Table 1.

Table 1. Summary of data sensemaking and data reasoning processes.

	Processes	Examples	Key References
Data Sensemaking	Summarization	Product of perceptual and cognitive mechanisms Implicit grouping of numbers that yields summary values (e.g., mean, variance)	[25,27,28,41–43]
Data Reasoning	Detecting Patterns	Detecting covariation between variables	[3,11,44,45]
	Detecting Differences	Noticing differences between sets	[28,41–43,46]
Scientific Data Reasoning	External Representation	External representations	[21,47,48]
	Scientific hypothesis testing	Conducting unconfounded experiments	[38,49–52]
	Probabilistic Conclusions	Evaluating the likelihood of conclusions or inferences	[4,20,53]
Limits to Data sensemaking and reasoning	Heuristics and biases	Confirmation bias, Anchoring effect	[15,32,54,55]
Sources of Change	Strategies	Acquiring better strategies for summarization, reasoning	[41,44,56,57]
	Instruction	Using data sensemaking to support formal reasoning and analysis	[58–60]

2. Data Sensemaking

We begin with a discussion of how data sensemaking occurs. As we will detail below, data sensemaking is the summarization of numerical information, a product of perceptual and cognitive mechanisms that summarize large quantities of information [25]. Numbers have unique properties such as relative magnitudes that are represented in both an approximate and exact fashion. Even young children have some elements of number sense that allows them to detect differences between quantities [61]. When these summarization mechanisms operate on number sets, they yield approximate representations of a set's statistical properties [25,28]. The following sections will outline the evidence for this account of data sensemaking, how it allows for the extraction of central tendency and variability, and how it changes over the course of development.

2.1. Sensemaking of Set Means and Variance

When people see a set of numerical data, they can summarize the data without using calculations, yielding approximate set properties such as means and variance [62]. Decades of evidence demonstrate that the properties of number sets can be summarized quickly [26,63] and accurately [43,64]. Without computation, people can detect and generate approximate means [41,43,65–70], detect relative variance [28,69,71], and increase their confidence in conclusions from larger samples as compared to smaller ones [42,43,68,69,71–74]. In one example of this early work with adults, participants were given a series of index cards each with a two-digit number, and asked to generate “a single value that best represented the series presented” [69] (p. 318). Participants generated a value that deviated from the arithmetic mean by less than 1%. When asked to estimate the mean from a set of numbers, and explicitly instructed not to calculate, participants were surprisingly accurate in generating an approximate mean (within ~3% of actual mean; [65]).

How might this process occur? We suggest that children and adults quickly summarize the properties of number sets similarly to how they summarize other types of complex information in their environments. This research tradition includes work from Gestalt psychologists [75] and recent research on ensemble perception and cognition [25,27]. Number sets may be summarized “automatically” [76] in that this may occur before conscious processing occurs (less than 100 MS; [77]) and even when instructions prohibit such processing [78]. Finally, reaction times are faster with larger sets than smaller sets with no loss in accuracy [76,79], further suggesting that summaries are the result of an automatic process.

Young children can summarize complex perceptual information, even in infancy [80]. However, summarization becomes more precise over the course of development [41,81]. For example, 6-month-olds can distinguish sets of dots with a 2:1 ratio (e.g., 10 from 20; [82]), while 6-year-olds can distinguish sets of dots at a 6:7 ratio [83]. Children as young as six can summarize the average happiness of a set of faces, but their summaries are less accurate than those of adolescents [84]. Given a set of objects (e.g., oranges), 4-year-olds can summarize the average size, though not as accurately as adults [81]. These findings suggest that summarization abilities emerge early in development and become refined over time.

Another numerical set characteristic is variability. The critical role of variability in empirical investigations has been noted for decades (e.g., [85]) and recently, Lehrer et al. [12] argued that variability is one of the fundamental issues it is necessary for students to understand when reasoning effectively about science. Functionally, it is only possible to measure the variability of a set of data, not a single data point; there must be data that vary to measure variability [86]. In considering what sets statistical reasoning apart from mathematical reasoning, Cobb and Moore [87] argue that although mathematical principles underlie many parts of statistical reasoning, it is the presence of variability due to real-world context that makes it statistical.

By first grade, children show an understanding that different variables are likely to differ in their variability when examining a data set [73], suggesting a conceptual understanding of underlying reasons to expect variation. Similarly, lay adults can demonstrate the ability to use variability in comparing data sets when the data are contextualized within a story, suggesting the likelihood of variability or not [71]. These findings indicate an expectation of variation when taking measurements from a heterogeneous sample, suggesting an understanding that variation is common in many contexts.

One component of understanding variability involves understanding the value of repeated measurements; without repeated measurements, there is nothing that can vary. Surveying 11-, 13-, and 15-year-olds indicated many areas of both clarity and confusion about experimental error, as well as the value of repeated measurements [88]. Although most students believed it was necessary to repeat measurements in science experiments, approximately 40% of participants in each age group focused solely on the means, and said the data with different variance levels were equally reliable because of the same average value, ignoring the variance. Detecting variability is also closely related to children's emerging understanding of the sources of this variability. Children understand by about age eight that measurement error is possible, and that repeated measurements, therefore, might not yield precisely the same results [89]. Children, especially 8-year-olds, were still not that likely to refer to measurement error in justifying their reasoning.

Other work has demonstrated that children and adults respond to variability information differently when they expect variability based on the context. For example, children ages 10–14 had an easier time using data to revise a belief from a noncausal one to a causal one [90]; the key complication in reasoning about noncausal links was in understanding measurement error and the value of repeated measurements to improve accuracy of estimations about data sets. In a converging set of findings, children ages 10–11 who expected a pattern of results indicating differences between conditions (such as in whether the length of a string affects the speed of a pendulum) were able to differentiate the small variance of noise from the larger differences between conditions [91]. At the same time, these children struggled more when they expected a difference but there was no true effect, and only small differences due to repeated measures. This point also emphasizes the close connection between data reasoning and scientific reasoning. For example, correctly interpreting data might hinge on recognizing the possibility of measurement error.

Sample size is also linked with reasoning about variability in data. There is a lot of evidence that people are more confident with larger samples of data (e.g., [68,69]). Further, many studies have found an interaction between sample size and variability when both are manipulated within the same study (e.g., [28,68,71,73]). For example, Jacobs and Narloch report that when samples had low variability, participants did not differentiate between

samples of 3 and 30, whereas in high-variability samples, even 7-year-olds responded differently based on sample size, and there were no age differences in the use of sample size. At the same time, there is evidence of failure to use sample size consistently in some contexts (e.g., [42,71]). More recent work has tried to reconcile apparently contradictory work about people's ability to use sample size in data reasoning, arguing that in fact weighted sample size follows a curvilinear function [74]. That is, with small sample sizes, participants are sensitive to sample size differences, but with large sample sizes, participants are no longer as sensitive to such differences and weight the differences much less. This finding suggests that numerical representations can affect broader data reasoning skills. Further, Obrecht found that intuitive number sense was also linked with the use of sample size, suggesting this factor may also play a role [74].

In addition to studies of implicit reasoning about variance, there are also several studies that have demonstrated that when children are asked to collect or are given data and asked to develop their own ways of summarizing the information, they can develop measures of center and variability that make sense to them. Additional design studies have focused on the integration of variation into describing data. For example, in figuring out how to display plant growth over two weeks of measurements, students had to consider how to represent both center (averaging) and variation [13]. Similarly, when children measure data in different contexts (for example, measuring the height of a flagpole with a handmade "height-o-meter" as compared to a standardized measuring tool), they observe a different amount of spread [14]. Another study involved asking 11-year-old children to each measure the perimeter of a table with a 15 cm ruler [92]. As expected, students' measurements varied, and then students worked in pairs to consider how to represent the full set of classroom data. These classroom studies also demonstrated a critical role for discussion as a means of advancing reasoning through relevant concepts to improve understanding.

2.2. Refining Data Sensemaking

What changes throughout development to refine this ability to summarize numerical data to estimate central tendency and variability? One contributing factor is acquiring and using more efficacious strategies (e.g., [93,94]). Children asked to summarize the spatial center in a series of dots used more strategies than adults, suggesting a less efficient process, and many of the strategies children used were not efficacious, resulting in fewer correct responses when compared to adults [41]. This result suggests that children's approaches to attending to and encoding information influence the resulting summaries [27]. Alibali et al. [56] recently proposed considering that the process of developing new strategies may be similar to a diathesis-stress model, in which there is an interaction between a "vulnerability to change . . . and a 'trigger' that actually provokes change (p. 164)". In other words, they suggest that once children have reached a point of being able to encode target problems in a way that makes key features salient, then it is possible that external input, such as feedback from successfully trying a new approach, will lead to the generation of new strategies. As they note, this does not fully explain the process of strategy generation, but it does suggest the importance of considering perceptual encoding as a factor in learning. There may also be value in considering more domain-specific models of change that occur within specific types of problems and across different age groups, for a more nuanced picture of the process [93].

As discussed above, adults are also adept at summarizing numerical information presented in sets [27]. Although people are often capable of summarizing data without conscious awareness, and encoding and drawing conclusions based on those summaries, one facet of learning to reason with data involves understanding what the summary values represent [95]. Students often gather or are given a series of individual data points, and then are asked to summarize the data. To do so effectively, they must recognize that reasoning about sets of data most commonly involves considering the data as an aggregate set, not as individual data points [96,97]. As students transition from informal reasoning about data to more scientific data reasoning, they are often taught formulas, enabling them for

example to compute means, and later standard deviations. However, the ability to apply formulas does not necessarily lead to understanding what the resulting values indicate, and how these summary values are related to the individual members of the set [35,98–100]. Nine-year-old children sometimes reason about a data set by referring only to a subset of the data [68]. Even university students who sometimes use aggregate reasoning are often inconsistent in their reasoning approaches and vary in whether they consider the full set of data or not, based on context [101,102].

Effective instruction makes use of a student's skills and prior knowledge to support their learning [103]. In the case of data reasoning, leveraging intuitions and prior knowledge about data can help students attend to relevant problem features [58], focus on possible strategies [59], and generate and attempt potential solutions that may be helpful in learning [60]. One instructional technique, preparation for learning, introduces students to relevant content before any formal instruction takes place [104]. In one application of this technique, students played a video game (Stats Invaders!) in which players identify the shape of the distributions in which invading aliens appear [58]. Students who played this video game before receiving instruction produced significantly higher scores at posttest than students who received instruction first, likely due to familiarizing students with statistical distributions before instruction began. Further exploration of how to bridge the gap between statistical intuitions and teaching statistical tools is important for clarifying this area. Statistical tools can augment and improve data reasoning, and provide some protection against cognitive biases. For example, statistical tools provide steps of formal analysis that control for sources of bias in informal analysis and allow for generalization beyond the data collected [4].

A different approach, productive failure, provides students with an opportunity to attempt to solve problems, and often fail, before instruction [59,60]. In two experiments comparing productive failure to direct instruction, students saw two instructional phases in one of two orders: (1) a data set with basketball performance and asked to determine the most consistent player and (2) direct instruction on calculating standard deviation [60]. Participants, who first explored the data and then were given direct instruction, outperformed students who were first given direct instruction before exploring. These findings suggest potential for broader applications of this concept.

To summarize, the evidence demonstrates that even young children can quickly summarize data resulting in approximate representations of statistical features such as variability, including the role of sample size. Much like summarization of sets of objects or other complex perceptual information, this process is rapid and occurs without any formal instruction. At the same time, variability is a more complex concept than the average, and children and adults often struggle to use variability information effectively. The following section will begin to explore how this initial data sensemaking underlies reasoning with data. Although children and adults can summarize large amounts of numerical information rapidly, drawing inferences and conclusions based on summary values may be skewed by mental shortcuts, known as heuristics.

2.3. Sensemaking and Reasoning from Associations between Variables

The section above described the initial process of data sensemaking that allows children and adults to summarize data. This process spares limited processing resources and provides information not available in individual numbers within a set. Summary values are one piece of information used to draw inferences. The following sections review research on both data sensemaking and reasoning from data. We combine these sections because detecting patterns in data or comparing data include both summarization and making sense of the patterns or differences that emerge from these summaries, and most of the tasks cited ask participants to reason with the data. Covariation refers to the relation between two or more variables and is one of the foundational principles in statistics and research [105]. Thus, one common application of data sensemaking and reasoning is within the contexts of reasoning about covariation between variables [106]. In the section below, we review

the experimental evidence, from research with children and adults, that illustrates data sensemaking and reasoning with covariation data and how strategy use influences informal data reasoning.

2.4. Making Sense of and Reasoning with Covariation Data

Data sensemaking often occurs when reasoning about covariation data, and drawing inferences from the patterns and relations between variables. Children and adults can detect differences in covariation data when those differences are large [44,45] or when covariation is presented within a constrained context [11]. More nuanced detection occurs as children acquire more sophisticated strategies for making sense of and interpreting covariation [44,57]. Early work in this vein [3] indicated children struggle with using covariation evidence to draw conclusions in line with the data, at least until ages 11 or 12. In many cases, children and even some adults referred to prior beliefs as justification, rather than the covariation evidence provided. For example, even if the data indicated more colds with carrot cake than chocolate cake (or no relationship), some children talked about how chocolate cake had more sugar, and was, therefore, less healthy and would lead to more colds. These findings have been used to suggest that children struggle with understanding covariation evidence, and have difficulty reasoning with this type of data.

However, follow-up work suggested that in fact young children could reason with covariation evidence when the tasks were simplified. For example, when given a less complex task, children by age six demonstrate an understanding of how covariation works. That is, they can use patterns of evidence to draw conclusions [11], particularly when the examples used tested equally plausible hypotheses. Similarly, young children ages 4–6 show evidence of the ability to use covariation evidence in drawing conclusions [107]. This suggests that young children can make sense of covariation data, when the differences are large.

Shaklee and colleagues report a series of studies in which they explored how people interpret covariation data (i.e., use strategies to reason with data) presented in contingency tables, in which there are four cells [29,44,45,108,109]. Participants were asked to consider whether there was a relationship between two dichotomous variables, such as the presence or absence of plant food and plant health or sickness. These studies demonstrated that children struggle to reason about contingency tables using sophisticated strategies, often ignoring some of the data. For example, Shaklee and Mims (1981) found that although strategy sophistication improved with age from fourth graders to college students, it was still only a minority of students even at the college level who used the most sophisticated strategy of conditional probability. Additional studies have found similar difficulties with strategy use in both children [29,110] and adults [57,109,111].

Taken together, the covariation results described above are consistent with data sensemaking that involves rapid summarization of data. In this case, detecting associations between variables would be possible with a mechanism that represents the event itself and represents an aggregate of multiple events of the same type. For example, seeing multiple instances of carrot eaters catching a cold would provide a strong pattern that should be readily detected by tracking cases [112]. However, only tracking this one outcome will lead to incorrect reasoning when there is a larger proportion of cases of carrot eaters who do not catch a cold. Finally, the consistency of the data (i.e., the strength of the correlation between variables) will make identifying the relations easier because more data points will be predictably in line with previous data points.

In many covariation studies, such as those described above, covariation is considered sufficient to demonstrate causation, and a mechanism linking two variables is not necessary. For example, in asking children to draw conclusions about the link between types of food and a cold, children are given no reason to believe one type of food would be better than another, beyond their knowledge of which foods are healthier than others. Similarly, figuring out which objects make a machine light up is determined by covariation evidence and temporal precedence. However, although covariation is one required piece of evidence

for inferring a causal relationship, it is not sufficient on its own. Further, analyzing data independent of theory is not what real-world scientists do, and there is an argument that it is important to consider the data in the context of one's prior knowledge about mechanisms that might link a cause and effect, enabling one to make an inference to the best explanation [37]. In many of these covariation studies, children are expected to ignore prior beliefs, even when prior beliefs suggest the data presented are implausible [2]. When given a potential explanatory mechanism, both children and adults reason based at least in part on these prior beliefs and mechanistic explanations, instead of exclusively on the data [3].

2.5. Sensemaking of and Reasoning with Group Comparisons

Another common inferential goal of scientific data reasoning is to determine if two (or more) groups are different on some outcome measure, and, again, data sensemaking and reasoning play a role. In this case, we are specifically talking about comparing categorical groups on a numerical or scale outcome measure. The origin of the first formal statistical test for group comparisons, Student's *t*-test, was to provide a method to compare samples of ingredients during beermaking [113]. People with training in statistics would typically use a *t*-test in making inferences about differences in an outcome between two groups. However, a small series of studies has demonstrated that children and adults often use the same components that are part of a formal *t*-test (i.e., differences between means, variance, and sample size) in drawing conclusions, even when comparing datasets without any calculation.

When comparing datasets, people generally rely most heavily on differences between means, with less attention to variance or sample size [35,68]. In a more recent study with adults, with more systematic manipulation of the mean difference and variance, larger mean differences and smaller variance in the datasets led to more accurate reasoning, more confidence in answers, and fewer visual fixations on the data [28]. These patterns suggest people summarize the data and compare the summaries quickly and accurately, without explicit computation. Other work has provided converging evidence that the magnitude of inferred (not computed) averages when comparing groups can depend in part on the magnitude of the values sampled [114], and that the ratio of means is a critical factor in reasoning about numerical data, such that accuracy of numerical perception varied in accordance with Weber's law (e.g., [83,115–117]).

A study of similar concepts looked at college students who compared pairs of consumer products in which the mean product ratings, the sample size, and the variance were all systematically manipulated. Participants focused most heavily on product ratings (magnitude of the outcome variable and the difference between means), and gave less weight to sample size. They gave the least weight of all to the sample variance [42].

Additional work has examined how college students compute analyses of variance (ANOVAs) intuitively, in which they are comparing four columns of data [46]. The data varied in their within-group variance and between-group variance, though students only saw raw data and this variance was not summarized. These students, similar to others described above, focused more in between group differences than within-group variance at the beginning of a semester-long statistics course.

The results reviewed in this section suggest that children and adults rely on data sensemaking; they make group comparisons quickly and without evidence for formal calculations, even in those who have received some formal instruction on data. Children and adults focus on differences between groups, as demonstrated by performance related to the statistical properties of the stimuli. This result pattern is consistent with a process of rapid summaries of both sets for comparisons [25,28]. In these models, individual numbers are represented as activation functions on a mental number line [118]. Multiple numbers are summarized by a secondary activation that is heightened with overlap among the individual values. The larger the distance between secondary activations, the faster the detection of difference. However, as with the detection of covariation, while children and adults are able to detect large differences, they become less accurate given smaller

differences (e.g., 9:10 ratio of means [28]) may attend to less diagnostic data features, and this informal detection may be influenced by cognitive biases.

3. Scientific Data Reasoning

Scientific reasoning processes provide tools to increase the validity and reliability of data reasoning while helping people reduce, or even avoid, common reasoning biases (e.g., confirmation bias) [119]. We will briefly describe three such tools that can improve data reasoning: external representations, scientific (i.e., theory-driven) hypothesis testing, and probabilistic conclusions. We recognize these descriptions are not comprehensive, but highlight a few key points about each topic.

3.1. External Representations

External representations refer to representations outside the mind that can be detected and processed by the perceptual and cognitive systems [21]. Examples of external representations related to numerical data are scatterplots, bar graphs, and columns of numbers in a spreadsheet. External representations allow us to record and display much larger amounts of data than can be held with fidelity in human memory [48]. Since internal representations are bound by the constraints of the human cognitive architecture, people can only attend to and process a finite amount of data at any given time [31].

External representations reduce this load by providing a representation of information, thereby allowing limited resources to be focused away from low-level process such as maintaining information in memory, to higher-order processes such as problem solving and reasoning. For example, 2nd graders were more likely to change beliefs in response to a diagram than an explanation [120], and 5th and 8th grade children were more successful in testing links between switch settings to make an electric train run when they kept external records [121]. A similar pattern holds with older participants: novice undergraduate and graduate students were more successful in solving a medical diagnostic problem when they created external representations of the problem (e.g., lists of symptoms or decision trees) than students who did not [47].

In addition to providing a reliable and durable record of data, external representations are accessible to others, and can allow for the discovery of patterns and higher-order features that would be difficult to detect in other formats (e.g., trends in a scatterplot; [21]). For example, scientists often compare their internal (i.e., mental) representations to external representations when reasoning with and interpreting ambiguous data [122].

However, even with external representations, cognitive biases can still influence data reasoning. For example, there is a tendency to underestimate means in bar graphs (though less so in point graphs), even in the presence of outliers [123]. This is likely because, as most models of graph comprehension suggest, people initially, and rapidly, summarize the main features of the graph, which forms the basis for subsequent inferences [124,125]. In sum, external representations provide powerful tools that aid with scientific data reasoning by reducing working memory burdens and making patterns and relations between variables more apparent; however, they are also subject to cognitive biases (e.g., mean underestimation).

3.2. Scientific Hypothesis Testing

Another tool for scientific data reasoning is scientific hypothesis testing. It has been documented that even young children engage in hypothesis testing [22]. The evidence for hypothesis testing and its development provides several seemingly contradictory findings. Developmental research demonstrates that young children have many of the rudiments of scientific reasoning [10,51,126,127]. When kindergartners use an inquiry-guided process, they can develop scientific questions and hypotheses more effectively than similar children not given such guidance [128]. There is also evidence that young children were more likely to seek information when evidence was inconsistent with their prior beliefs than when evidence was consistent with their beliefs [49]. Children as young as five spontaneously performed contrastive tests (i.e., compared two different experimental setups), in which

they tested whether a machine lit up with or without a targeted variable [51]. These findings collectively suggest more robust scientific reasoning ability in children than assumed in early developmental research [129].

At the same time, evidence from several studies suggest that children and adults often fail to conduct unconfounded hypothesis tests, as would occur in scientific experimentation [40,51,130]. Children often conduct confounded experiments before they have received instruction on this topic [40] and sometimes struggle to construct unconfounded hypotheses in unconstrained contexts such as discovery learning [50]. Adults sometimes do not perform contrastive tests and sometimes fail to identify causal variables [131]. There is research that demonstrates a tendency for children [132] and adults [54] to seek to confirm beliefs. Additionally, preschoolers sometimes do not seek disconfirming evidence after hearing misleading testimonial evidence [133]. This pattern of results might arise from either lacking knowledge about scientific hypothesis testing or not implementing this knowledge correctly. Further, even when seeking evidence, children and adults sometimes misinterpret or misperceive data such that new data conform with their prior beliefs, despite the misconceptions of those prior beliefs [15].

This pattern of evidence likely suggests a developmental and educational trajectory in which children's curiosity drives them to understand the world by seeking information [49,51,134,135]. Children quickly acquire impressive skills for information seeking [49] and evidence evaluation [10]. However, these skills are limited by children's emerging understanding of scientific experimentation [50], implementing this knowledge in novel contexts [38,136], and cognitive biases (e.g., confirmation bias [15,119]). The acquisition and use of scientific reasoning skills improves how people evaluate and understand the data about which they are reasoning, which improves the quality of the conclusions drawn from data. In short, the acquisition and application of scientific hypothesis testing can help to protect reasoners from errors that may reduce the accuracy of their data [52].

3.3. Probabilistic Conclusions

Data reasoning leads to conclusions, but these conclusions are always probabilistic rather than deterministic (e.g., [4,53]). Science education, including scientific data reasoning, often presents scientific conclusions as definitive [20]. Including acknowledgement of the uncertainty inherent in scientific data is an important, but often overlooked, area of science education (e.g., [137–139]). When children work with real-world data, with its variability and uncertainty, they often come to understand the nature of science more effectively [140]. Young children often appear to have a bias towards deterministic conclusions, preferring to select a single outcome when multiple outcomes are possible [141,142]. This tendency to look for a single conclusion is robust but is reduced with age [142] and can be reduced after multiple training experiences [9].

At the same time, Denison and Xu [143] argue that the majority of empirical evidence into infant (and primate) reasoning under certainty suggests that they use probabilistic reasoning in drawing conclusions. Young children have some intuitions about probability that help them make sense of situations such as the likelihood of selecting a specific object from a target set. In one recent experiment, 6- and 7-year-old children were shown a set of white and red balls with specific ratios of difference between red and white balls (e.g., 1.10–9.90; [144]). In line with previous results reported above, children's accuracy in selecting the most likely ball was closely associated with the ratio of difference.

These intuitions of data sensemaking can influence reasoning from data, but the understanding that inferences from data must be probabilistic is necessary for effective scientific reasoning, despite its challenges (e.g., [12,53]). Even when adding inferential statistics to the toolkit, there can still be a wide range of approaches taken by experts in the field [145]. That variation is one of the many challenges in thinking through scientific data reasoning, and a factor that makes teaching these concepts especially difficult. How do we leverage intuitions about data to promote scientific data reasoning?

4. Heuristics in Data Reasoning

A key limitation of all data reasoning is that humans are subject to cognitive biases, and when reasoning with data, we can fall prey to them. Tversky and Kahneman's classic work on heuristics and biases [32] suggests several ways in which shortcuts we often take to reason about data can lead us astray. For example, people are more likely to think things that come to mind easily are more common than those that do not come to mind as quickly, a phenomenon known as the availability heuristic. People also often estimate magnitude by anchoring to an initial value. When seeing data, the anchor then affects conclusions, and adjustment for the anchor is often insufficient. Additionally, people often test hypotheses with a confirmation bias, looking for evidence to support their initial beliefs rather than seeking and evaluating evidence independent of hypotheses (e.g., [15,54,55]).

Although mental shortcuts can lead to suboptimal conclusions, under some conditions, shortcuts may lead to better conclusions and decisions than deliberative reasoning, a phenomenon termed adaptive heuristics [146]. For example, when selecting the best mutual fund for retirement investments, a simple, adaptive heuristic in which one allocates equally to all options, outperformed data-driven models that far exceeded human processing limits [147]. In this case, the use of a simple strategy was highly efficient and could easily be implemented within limited cognitive capacity. Adaptive heuristics are useful when thinking about reasoning with data because we often have to make sense of large amounts of information (e.g., data) and formal data calculations require significant time, energy, and working memory capacity [148].

However, reliance on heuristics alone might result in suboptimal conclusions, as described above. Recent evidence demonstrates that training in the scientific process leads to reduced susceptibility to cognitive biases [149]. It is important to note that heuristics are not supplanted by scientific reasoning. Heuristics continue to operate even for experts and may compete for cognitive resources [150]. Experts might use heuristics in a more controlled and deliberate fashion than novices [151]. In addition, reliance on prior knowledge about mechanisms, and assessing data in light of that knowledge, often makes sense in a scientific context (e.g., [2,37,152]). For example, if an initial analysis provides evidence against a well-established pattern of evidence, it is often reasonable to check the data and analysis or even replicate a study before abandoning one's hypothesis. Additionally, consideration of the plausibility of the proposed mechanisms for an effect play a role. In the following section, we will discuss how intuitive data reasoning strategies (e.g., heuristics) play a role in data reasoning and how these processes can be leveraged through instruction to help students learn scientific data reasoning.

5. Future Directions

Many basic research questions remain in this realm of data sensemaking, informal reasoning with data, and scientific data reasoning. For example, although the evidence presented above suggests rapid summarization of data sets (e.g., [26,29,68]), more research is needed to determine the extent to which summarizing data is made on the basis of the same mechanisms underlying ensemble perception and cognition. Further, as we have discussed, data reasoning occurs in a wide ranges of contexts, including scientific reasoning, science education, decision-making, and other fields. We have focused on scientific reasoning and a little bit of the science education literature. Further exploration of differences in data reasoning across disciplines, with and without the supports of external representations, scientific hypothesis testing, and probabilistic conclusions, would also help in understanding the process of data reasoning more thoroughly.

We have suggested that reasoning with data begins with data sensemaking, a rapid summarization process that reduces processing burdens while providing information about the statistical properties of number sets. This process appears to improve through development and education, resulting in more accurate summaries. We suggested that one important factor underlying these improvements is the acquisition and use of more effective strategies, which are developed with experience and education. Data reasoning

is drawing inferences from the data, along with prior knowledge and other relevant information. Much like data summaries, data reasoning is often “accurate enough” for everyday contexts. One limitation to accuracy for summaries and inferences is reasoning biases, such as the confirmation bias or the tendency to seek data consistent with prior expectations. We propose that the acquisition of scientific data reasoning provides tools that improve the fidelity of the data itself, the conditions through which data are acquired, representation of data (e.g., figures), and the types of conclusions drawn from data. An important future direction is to evaluate this model experimentally.

Our proposed model needs direct testing of its components, though focus on the concepts at a fine grain size provides researchers with opportunities to evaluate elements of the model or the model itself. Our predictions about relatively automatic summarization of data sets can be evaluated directly. Educators can implement parts of our model individually or in concert. For example, lessons on data interpretation can encourage reliance on data summarization, with instruction guiding students to describe patterns and compare data sets in consistent ways. Below, we highlight several specific suggestions for future directions.

One complex topic in need of much further exploration is the interplay between prior knowledge and data reasoning. Although there is a fair amount of work about integrating theory and evidence (e.g., [2,3,90,152,153]), there is less work on how prior beliefs interact with different types of numerical data (e.g., [13,14,91]). There is evidence that prior knowledge increases attention to diagnostic features [154] and helps reasoners solve problems more effectively [155]. However, this attention to diagnostic features has not to our knowledge been tested with data reasoning.

In addition, there are many educational applications of data reasoning, and specifically scientific data reasoning. Future research aimed at effective application of these concepts in the classroom can be beneficial both to understanding scientific data reasoning, and to developing best practices in education. As discussed above, classroom studies in which students develop their own measures of description and inferences from data have been shown to facilitate a more comprehensive understanding of concepts such as the aggregation of data and variability, building on initial intuitions (e.g., [13,14,35,156]). Considered within the framing outlined above by Alibali et al. [93], the classroom conversations could be considered a potential trigger for provoking changes in strategies used to approach these problems, and in turn, increase learning. This process can work in informal data reasoning or scientific data reasoning contexts. Follow-up studies could directly examine strategy acquisition and be used to develop a more comprehensive understanding of how strategies aid in learning about data reasoning.

There is a lot of work demonstrating the efficacy of classroom interventions or curricular approaches in improving people’s ability to reason statistically [96]. A meta-analysis of scientific reasoning interventions, targeting a wider range of topics than just data reasoning, indicated there was a small effect in classroom interventions across ages [19]. Similarly, there are many demonstrations of efficacy of specific tools aiding in data reasoning within lab contexts (e.g., [56,120,157,158]). However, scaling up these interventions into more effective curricula at all levels (including teacher training) remains a challenge.

Teaching materials are also important in facilitating (or unintentionally hindering) student learning. Textbooks play an important role in student learning, and limitations in textbook content can affect student learning. Children can acquire misconceptions through misaligned instructional materials. One source of misconception can be examples in textbooks. A notable example from mathematics is children’s misconception of the equal sign, in which children interpret the equal sign as a signal for executing an operation rather than balancing both sides of an equation [159]. An analysis of math textbooks demonstrated that most practice problems had the same structure (e.g., $3 + 5 = ?$) that is consistent with this misconception [160]. Another study of middle school science textbooks showed they typically include limited guidance in appropriate use of data [161]. In fact, the majority of data reasoning activities in science texts provided little guidance on how to analyze or draw inferences from data formally. Thus, one step that can help improve student learning

of data concepts involves improved integration of descriptions and applied exercises in textbooks used in science classes. This research demonstrates the importance of using instructional materials that do not promote biases or misconceptions [160,162].

Finally, one last suggested future direction is investigating the role of intuitions and potential misconceptions, both about science and about data, in scientific data reasoning. One difficulty in science education is that many scientific phenomena are challenging to understand, and in many cases intuitions conflict with scientific consensus, such as in understanding of physical principles of heat and motion as well as biological principles of inheritance and evolution [163]. Thus, although intuitions about data can be useful in data reasoning, intuitions about conceptual content sometimes lead people to incorrect beliefs and misconceptions. Indeed, there is some evidence it can be at the observation stage where incorrect prior beliefs interfere with accurate perception of physical phenomena and the gathering of potentially-informative data [15].

6. Conclusions

We proposed a model of data reasoning and its relation to scientific reasoning. Specifically, we suggest that data reasoning begins developmentally with data sensemaking, a relatively automatic process rooted in perceptual mechanisms that summarize large quantities of information in the environment. As these summarization mechanisms operate on number sets, they yield approximate representations of statistical properties, such as central tendency and variation. This information is then available for drawing inferences from data. However, both data sensemaking and informal data reasoning may lead to erroneous conclusions due to cognitive biases or heuristics. The acquisition of scientific data reasoning helps to reduce these biases by providing tools and procedures that improve data reasoning. These tools include external representations, scientific hypothesis testing, and drawing probabilistic conclusions. Although data sensemaking and informal data reasoning are not supplanted by scientific data reasoning, these skills can be leveraged to improve learning of science and reasoning with data.

Author Contributions: The authors contributed equally to all phases of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

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

References

1. Klahr, D.; Dunbar, K. Dual space search during scientific reasoning. *Cogn. Sci.* **1988**, *12*, 1–48. [CrossRef]
2. Koslowski, B. *Theory and Evidence: The Development of Scientific Reasoning*; MIT Press: Cambridge, MA, USA, 1996; ISBN 978-0-262-11209-4.
3. Kuhn, D.; Amsel, E.; O’Loughlin, M.; Schauble, L.; Leadbeater, B.; Yotive, W. *The Development of Scientific Thinking Skills*; Academic Press: San Diego, CA, USA, 1988; ISBN 978-0-12-428430-2.
4. Grolemond, G.; Wickham, H. A cognitive interpretation of data analysis. *Int. Stat. Rev.* **2014**, *82*, 184–204. [CrossRef]
5. Zimmerman, C.; Klahr, D. Development of scientific thinking. In *Stevens’ Handbook of Experimental Psychology and Cognitive Neuroscience*; Wixted, J.T., Ed.; John Wiley & Sons, Inc.: New York, NY, USA, 2018; pp. 1–25, ISBN 978-1-119-17016-7.
6. Fischer, F.; Kollar, I.; Ufer, S.; Sodian, B.; Hussmann, H.; Pekrun, R.; Neuhaus, B.; Dorner, B.; Pankofer, S.; Fischer, M.; et al. Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learn. Res.* **2014**, *2*, 28–45.
7. Morris, B.J.; Croker, S.; Masnick, A.M.; Zimmerman, C. The emergence of scientific reasoning. In *Current Topics in Children’s Learning and Cognition*; Kloos, H., Ed.; InTech: Rijeka, Croatia, 2012; ISBN 978-953-51-0855-9.
8. Zimmerman, C. The development of scientific reasoning skills. *Dev. Rev.* **2000**, *20*, 99–149. [CrossRef]
9. Klahr, D.; Chen, Z. Overcoming the positive-capture strategy in young children: Learning about indeterminacy. *Child Dev.* **2003**, *74*, 1275–1296. [CrossRef]
10. Sodian, B.; Zaitchik, D.; Carey, S. Young children’s differentiation of hypothetical beliefs from evidence. *Child Dev.* **1991**, *62*, 753–766. [CrossRef]

11. Ruffman, T.; Perner, J.; Olson, D.R.; Doherty, M. Reflecting on scientific thinking: Children's understanding of the hypothesis-evidence relation. *Child Dev.* **1993**, *64*, 1617–1636. [CrossRef]
12. Lehrer, R.; Schauble, L.; Wisittanawat, P. Getting a grip on variability. *Bull. Math. Biol.* **2020**, *82*, 106. [CrossRef]
13. Lehrer, R.; Schauble, L. Modeling natural variation through distribution. *Am. Educ. Res. J.* **2004**, *41*, 635–679. [CrossRef]
14. Petrosino, A.J.; Lehrer, R.; Schauble, L. Structuring error and experimental variation as distribution in the fourth grade. *Math. Think. Learn.* **2003**, *5*, 131–156. [CrossRef]
15. Chinn, C.A.; Malhotra, B.A. Children's responses to anomalous scientific data: How is conceptual change impeded? *J. Educ. Psychol.* **2002**, *94*, 327–343. [CrossRef]
16. Koslowski, B.; Okagaki, L.; Lorenz, C.; Umbach, D. When covariation is not enough: The role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Dev.* **1989**, *60*, 1316–1327. [CrossRef]
17. Kuhn, D. What is scientific thinking and how does it develop? In *The Wiley-Blackwell Handbook of Childhood Cognitive Development*, 2nd ed.; Wiley-Blackwell: Malden, MA, USA, 2011; pp. 497–523, ISBN 978-1-4051-9116-6.
18. van Lieshout, L.L.; de Lange, F.P.; Cools, R. Why so curious? Quantifying mechanisms of information seeking. *Curr. Opin. Behav. Sci.* **2020**, *35*, 112–117. [CrossRef]
19. Engelmann, K.; Neuhaus, B.J.; Fischer, F. Fostering scientific reasoning in education—Meta-analytic evidence from intervention studies. *Educ. Res. Eval.* **2016**, *22*, 333–349. [CrossRef]
20. Shah, P.; Michal, A.; Ibrahim, A.; Rhodes, R.; Rodriguez, F. What makes everyday scientific reasoning so challenging? In *Psychology of Learning and Motivation*; Elsevier: Amsterdam, The Netherlands, 2017; Volume 66, pp. 251–299, ISBN 978-0-12-812118-4.
21. Zhang, J. The nature of external representations in problem solving. *Cogn. Sci.* **1997**, *21*, 179–217. [CrossRef]
22. Koerber, S.; Mayer, D.; Osterhaus, C.; Schwippert, K.; Sodian, B. The development of scientific thinking in elementary school: A comprehensive inventory. *Child Dev.* **2015**, *86*, 327–336. [CrossRef]
23. Di Blas, N.; Mazuran, M.; Paolini, P.; Quintarelli, E.; Tanca, L. Exploratory computing: A comprehensive approach to data sensemaking. *Int. J. Data Sci. Anal.* **2017**, *3*, 61–77. [CrossRef]
24. Koesten, L.; Gregory, K.; Groth, P.; Simperl, E. Talking datasets—Understanding data sensemaking behaviours. *Int. J. Hum.-Comput. Stud.* **2021**, *146*, 102562. [CrossRef]
25. Alvarez, G.A. Representing multiple objects as an ensemble enhances visual cognition. *Trends Cogn. Sci.* **2011**, *15*, 122–131. [CrossRef] [PubMed]
26. Brezis, N.; Bronfman, Z.Z.; Usher, M. Adaptive spontaneous transitions between two mechanisms of numerical averaging. *Sci. Rep.* **2015**, *5*, 10415. [CrossRef]
27. Whitney, D.; Yamanashi Leib, A. Ensemble perception. *Annu. Rev. Psychol.* **2018**, *69*, 105–129. [CrossRef] [PubMed]
28. Morris, B.J.; Masnick, A.M. Comparing data sets: Implicit summaries of the statistical properties of number sets. *Cogn. Sci.* **2015**, *39*, 156–170. [CrossRef] [PubMed]
29. Shaklee, H.; Paszek, D. Covariation judgment: Systematic rule use in middle childhood. *Child Dev.* **1985**, *56*, 1229–1240. [CrossRef]
30. Makar, K.; Rubin, A. Learning about statistical inference. In *International Handbook of Research in Statistics Education*; Ben-Zvi, D., Makar, K., Garfield, J., Eds.; Springer International Handbooks of Education; Springer International Publishing: Cham, Switzerland, 2018; pp. 261–294, ISBN 978-3-319-66195-7.
31. Cowan, N. Mental objects in working memory: Development of basic capacity or of cognitive completion? *Adv. Child Dev. Behav.* **2017**, *52*, 81–104.
32. Tversky, A.; Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science* **1974**, *185*, 1124–1131. [CrossRef]
33. Makar, K. Developing young children's emergent inferential practices in statistics. *Math. Think. Learn.* **2016**, *18*, 1–24. [CrossRef]
34. Makar, K.; Bakker, A.; Ben-Zvi, D. The reasoning behind informal statistical inference. *Math. Think. Learn.* **2011**, *13*, 152–173. [CrossRef]
35. Watson, J.M.; Moritz, J.B. The beginning of statistical inference: Comparing two data sets. *Educ. Stud. Math.* **1999**, *37*, 145–168. [CrossRef]
36. Zieffler, A.; Garfield, J.; Delmas, R.; Reading, C. A framework to support research on informal inferential reasoning. *Stat. Educ. Res. J.* **2008**, *7*, 40–58.
37. Koslowski, B. Inference to the best explanation (IBE) and the causal and scientific reasoning of nonscientists. In *Psychology of Science: Implicit and Explicit Processes*; Proctor, R.W., Capaldi, E.J., Eds.; OUP: New York, NY, USA, 2012; ISBN 978-0-19-975362-8.
38. Chen, Z.; Klahr, D. All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Dev.* **1999**, *70*, 1098–1120. [CrossRef]
39. Krüger, D.; Hartmann, S.; Nordmeier, V.; Upmeyer zu Belzen, A. Measuring scientific reasoning competencies. In *Student Learning in German Higher Education: Innovative Measurement Approaches and Research Results*; Zlatkin-Troitschanskaia, O., Pant, H.A., Toepper, M., Lautenbach, C., Eds.; Springer Fachmedien: Wiesbaden, Germany, 2020; pp. 261–280, ISBN 978-3-658-27886-1.
40. Klahr, D. *Exploring Science: The Cognition and Development of Discovery Processes*; MIT Press: Cambridge, MA, USA, 2002; ISBN 978-0-262-61176-3.
41. Jones, P.R.; Dekker, T.M. The development of perceptual averaging: Learning what to do, not just how to do it. *Dev. Sci.* **2018**, *21*, e12584. [CrossRef] [PubMed]
42. Obrecht, N.A.; Chapman, G.B.; Gelman, R. Intuitive t tests: Lay use of statistical information. *Psychon. Bull. Rev.* **2007**, *14*, 1147–1152. [CrossRef] [PubMed]

43. Peterson, C.R.; Beach, L.R. Man as an intuitive statistician. *Psychol. Bull.* **1967**, *68*, 29–46. [CrossRef]
44. Shaklee, H.; Mims, M. Development of rule use in judgments of covariation between events. *Child Dev.* **1981**, *52*, 317–325. [CrossRef]
45. Shaklee, H.; Holt, P.; Elek, S.; Hall, L. Covariation judgment: Improving rule use among children, adolescents, and adults. *Child Dev.* **1988**, *59*, 755–768. [CrossRef]
46. Trumppower, D.L. Formative use of intuitive Analysis of Variance. *Math. Think. Learn.* **2013**, *15*, 291–313. [CrossRef]
47. Martin, L.; Schwartz, D.L. Prospective adaptation in the use of external representations. *Cogn. Instr.* **2009**, *27*, 370–400. [CrossRef]
48. Zhang, J.; Wang, H. An exploration of the relations between external representations and working memory. *PLoS ONE* **2009**, *4*, e6513. [CrossRef]
49. Bonawitz, E.B.; van Schijndel, T.J.P.; Friel, D.; Schulz, L. Children balance theories and evidence in exploration, explanation, and learning. *Cogn. Psychol.* **2012**, *64*, 215–234. [CrossRef]
50. Klahr, D.; Nigam, M. The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychol. Sci.* **2004**, *15*, 661–667. [CrossRef]
51. Köksal-Tuncer, Ö.; Sodian, B. The development of scientific reasoning: Hypothesis testing and argumentation from evidence in young children. *Cogn. Dev.* **2018**, *48*, 135–145. [CrossRef]
52. Sandoval, W.A.; Sodian, B.; Koerber, S.; Wong, J. Developing children’s early competencies to engage with science. *Educ. Psychol.* **2014**, *49*, 139–152. [CrossRef]
53. Wild, C.J.; Utts, J.M.; Horton, N.J. What is statistics? In *International Handbook of Research in Statistics Education*; Ben-Zvi, D., Makar, K., Garfield, J., Eds.; Springer International Handbooks of Education; Springer International Publishing: Cham, Switzerland, 2018; pp. 5–36, ISBN 978-3-319-66195-7.
54. Klayman, J.; Ha, Y. Confirmation, disconfirmation, and information in hypothesis testing. *Psychol. Rev.* **1987**, *94*, 211–228. [CrossRef]
55. Mynatt, C.R.; Doherty, M.E.; Tweney, R.D. Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Q. J. Exp. Psychol.* **1977**, *29*, 85–95. [CrossRef]
56. Alibali, M.W.; Crooks, N.M.; McNeil, N.M. Perceptual support promotes strategy generation: Evidence from equation solving. *Br. J. Dev. Psychol.* **2018**, *36*, 153–168. [CrossRef]
57. Batanero, C.; Estepa, A.; Godino, J.D.; Green, D.R. Intuitive strategies and preconceptions about association in contingency tables. *J. Res. Math. Educ.* **1996**, *27*, 151–169. [CrossRef]
58. Arena, D.A.; Schwartz, D.L. Experience and explanation: Using videogames to prepare students for formal instruction in statistics. *J. Sci. Educ. Technol.* **2014**, *23*, 538–548. [CrossRef]
59. Kapur, M. Productive failure. *Cogn. Instr.* **2008**, *26*, 379–424. [CrossRef]
60. Kapur, M. Productive failure in learning math. *Cogn. Sci.* **2014**, *38*, 1008–1022. [CrossRef]
61. Dehaene, S. *The Number Sense: How the Mind Creates Mathematics*; Rev and Updated Edition; Oxford University Press: New York, NY, USA, 2011; ISBN 978-0-19-975387-1.
62. Hyde, D.C. Two systems of non-symbolic numerical cognition. *Front. Hum. Neurosci.* **2011**, *5*, 150. [CrossRef]
63. Rosenbaum, D.; de Gardelle, V.; Usher, M. Ensemble perception: Extracting the average of perceptual versus numerical stimuli. *Atten. Percept. Psychophys.* **2021**, *83*, 956–969. [CrossRef] [PubMed]
64. Malmi, R.A.; Samson, D.J. Intuitive averaging of categorized numerical stimuli. *J. Verbal Learn. Verbal Behav.* **1983**, *2*, 547–559. [CrossRef]
65. Beach, L.R.; Swenson, R.G. Intuitive estimation of means. *Psychon. Sci.* **1966**, *5*, 161–162. [CrossRef]
66. Grebstein, L.C. Relative accuracy of actuarial prediction, experienced clinicians, and graduate students in a clinical judgment task. *J. Consult. Psychol.* **1963**, *27*, 127–132. [CrossRef]
67. Irwin, F.W.; Smith, W.A.S. Value, cost, and information as determiners of decision. *J. Exp. Psychol.* **1957**, *54*, 229–232. [CrossRef] [PubMed]
68. Masnick, A.M.; Morris, B.J. Investigating the development of data evaluation: The role of data characteristics. *Child Dev.* **2008**, *79*, 1032–1048. [CrossRef]
69. Spencer, J. Estimating averages. *Ergonomics* **1961**, *4*, 317–328. [CrossRef]
70. Spencer, J. A further study of estimating averages. *Ergonomics* **1963**, *6*, 255–265. [CrossRef]
71. Obrecht, N.A.; Chapman, G.B.; Suárez, M.T. Laypeople do use sample variance: The effect of embedding data in a variance-implicating story. *Think. Reason.* **2010**, *16*, 26–44. [CrossRef]
72. Irwin, F.W.; Smith, W.A.S.; Mayfield, J.F. Tests of two theories of decision in an “expanded judgment” situation. *J. Exp. Psychol.* **1956**, *51*, 261–268. [CrossRef]
73. Jacobs, J.E.; Narloch, R.H. Children’s use of sample size and variability to make social inferences. *J. Appl. Dev. Psychol.* **2001**, *22*, 311–331. [CrossRef]
74. Obrecht, N.A. Sample size weighting follows a curvilinear function. *J. Exp. Psychol. Learn. Mem. Cogn.* **2019**, *45*, 614–626. [CrossRef]
75. Wagemans, J.; Elder, J.H.; Kubovy, M.; Palmer, S.E.; Peterson, M.A.; Singh, M.; von der Heydt, R. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychol. Bull.* **2012**, *138*, 1172–1217. [CrossRef] [PubMed]

76. Lee, K.R.; Sobel, K.V.; York, A.K.; Puri, A.M. Dissociating parallel and serial processing of numerical value. *J. Numer. Cogn.* **2018**, *4*, 360–379. [CrossRef]
77. Van Opstal, F.; de Lange, F.P.; Dehaene, S. Rapid parallel semantic processing of numbers without awareness. *Cognition* **2011**, *120*, 136–147. [CrossRef]
78. Dehaene, S.; Akhvein, R. Attention, automaticity, and levels of representation in number processing. *J. Exp. Psychol. Learn. Mem. Cogn.* **1995**, *21*, 314–326. [CrossRef]
79. Lee, K.R.; Dague, T.D.; Sobel, K.V.; Paternoster, N.J.; Puri, A.M. Set size and ensemble perception of numerical value. *Atten. Percept. Psychophys.* **2021**, *83*, 1169–1178. [CrossRef] [PubMed]
80. Zosh, J.M.; Halberda, J.; Feigenson, L. Memory for multiple visual ensembles in infancy. *J. Exp. Psychol. Gen.* **2011**, *140*, 141–158. [CrossRef]
81. Sweeny, T.D.; Wurnitsch, N.; Gopnik, A.; Whitney, D. Ensemble perception of size in 4–5-year-old children. *Dev. Sci.* **2015**, *18*, 556–568. [CrossRef] [PubMed]
82. Xu, F.; Spelke, E.S. Large number discrimination in 6-month-old infants. *Cognition* **2000**, *74*, B1–B11. [CrossRef]
83. Halberda, J.; Feigenson, L. Developmental change in the acuity of the “number sense”: The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Dev. Psychol.* **2008**, *44*, 1457–1465. [CrossRef]
84. Rhodes, G.; Neumann, M.; Ewing, L.; Bank, S.; Read, A.; Engfors, L.M.; Emiechel, R.; Palermo, R. Ensemble coding of faces occurs in children and develops dissociably from coding of individual faces. *Dev. Sci.* **2018**, *21*, e12540. [CrossRef] [PubMed]
85. Wild, C.J.; Pfannkuch, M. Statistical thinking in empirical enquiry. *Int. Stat. Rev.* **1999**, *67*, 223–248. [CrossRef]
86. Lee, V.R.; Wilkerson, M.H. Data use by middle and secondary students in the digital age: A status report and future prospects. In *Commissioned Paper for the National Academies of Sciences, Engineering, and Medicine, Board on Science Education, Committee on Science Investigations and Engineering Design for Grades 6–12*; National Academy of Sciences Engineering, and Medicine: Washington, DC, USA, 2018; p. 43.
87. Cobb, G.W.; Moore, D.S. Mathematics, statistics, and teaching. *Am. Math. Mon.* **1997**, *104*, 801–823. [CrossRef]
88. Lubben, F.; Millar, R. Children’s ideas about the reliability of experimental data. *Int. J. Sci. Educ.* **1996**, *18*, 955–968. [CrossRef]
89. Masnick, A.M.; Klahr, D. Error matters: An initial exploration of elementary school children’s understanding of experimental error. *J. Cogn. Dev.* **2003**, *4*, 67–98. [CrossRef]
90. Kanari, Z.; Millar, R. Reasoning from data: How students collect and interpret data in science investigations. *J. Res. Sci. Teach.* **2004**, *41*, 748–769. [CrossRef]
91. Masnick, A.M.; Klahr, D.; Knowles, E.R. Data-driven belief revision in children and adults. *J. Cogn. Dev.* **2017**, *18*, 87–109. [CrossRef]
92. Lehrer, R. Modeling signal-noise processes supports student construction of a hierarchical image of sample. *SERJ* **2017**, *16*, 64–85. [CrossRef]
93. Alibali, M.W.; Brown, S.A.; Menendez, D. Understanding strategy change: Contextual, individual, and metacognitive factors. *Adv. Child Dev. Behav.* **2019**, *56*, 227–256. [PubMed]
94. Siegler, R.; Jenkins, E.A. *How Children Discover New Strategies*; Psychology Press: New York, NY, USA, 1989; ISBN 978-1-315-80774-4.
95. Hancock, C.; Kaput, J.J.; Goldsmith, L.T. Authentic inquiry with data: Critical barriers to classroom implementation. *Educ. Psychol.* **1992**, *27*, 337–364. [CrossRef]
96. Garfield, J.; Ben-Zvi, D. How students learn statistics revisited: A current review of research on teaching and learning statistics. *Int. Stat. Rev.* **2007**, *75*, 372–396. [CrossRef]
97. Konold, C.; Higgins, T.; Russell, S.J.; Khalil, K. Data seen through different lenses. *Educ. Stud. Math.* **2015**, *88*, 305–325. [CrossRef]
98. Konold, C.; Pollatsek, A. Data analysis as the search for signals in noisy processes. *J. Res. Math. Educ.* **2002**, *33*, 259. [CrossRef]
99. Mokros, J.; Russell, S.J. Children’s concepts of average and representativeness. *J. Res. Math. Educ.* **1995**, *26*, 20–39. [CrossRef]
100. Pollatsek, A.; Lima, S.; Well, A.D. Concept or computation: Students’ understanding of the mean. *Educ. Stud. Math.* **1981**, *12*, 191–204. [CrossRef]
101. Buffler, A.; Allie, S.; Lubben, F. The development of first year physics students’ ideas about measurement in terms of point and set paradigms. *Int. J. Sci. Educ.* **2001**, *23*, 1137–1156. [CrossRef]
102. Lubben, F.; Campbell, B.; Buffler, A.; Allie, S. Point and set reasoning in practical science measurement by entering university freshmen. *Sci. Educ.* **2001**, *85*, 311–327. [CrossRef]
103. Willingham, D.T. *Why Don’t Students Like School? A Cognitive Scientist Answers Questions about How the Mind Works and What It Means for the Classroom*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2021; ISBN 978-1-119-71580-1.
104. Schwartz, D.L.; Martin, T. Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cogn. Instr.* **2004**, *22*, 129–184. [CrossRef]
105. Kerlinger, F.N.; Lee, H.B. *Foundations of Behavioral Research*; Harcourt College Publishers: Fort Worth, TX, USA, 2000; ISBN 978-0-15-507897-0.
106. Biehler, R.; Frischemeier, D.; Reading, C.; Shaughnessy, J.M. Reasoning about data. In *International Handbook of Research in Statistics Education*; Ben-Zvi, D., Makar, K., Garfield, J., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 139–192.
107. Koerber, S.; Sodian, B.; Thoermer, C.; Nett, U. Scientific reasoning in young children: Preschoolers’ ability to evaluate covariation evidence. *Swiss J. Psychol.* **2005**, *64*, 141–152. [CrossRef]

108. Shaklee, H.; Mims, M. Sources of error in judging event covariations: Effects of memory demands. *J. Exp. Psychol. Learn. Mem. Cogn.* **1982**, *8*, 208–224. [CrossRef]
109. Shaklee, H.; Tucker, D. A rule analysis of judgments of covariation between events. *Mem. Cogn.* **1980**, *8*, 459–467. [CrossRef]
110. Obersteiner, A.; Bernhard, M.; Reiss, K. Primary school children's strategies in solving contingency table problems: The role of intuition and inhibition. *ZDM Math. Educ.* **2015**, *47*, 825–836. [CrossRef]
111. Osterhaus, C.; Magee, J.; Saffran, A.; Alibali, M.W. Supporting successful interpretations of covariation data: Beneficial effects of variable symmetry and problem context. *Q. J. Exp. Psychol.* **2019**, *72*, 994–1004. [CrossRef]
112. Jung, Y.; Walther, D.B.; Finn, A.S. Children automatically abstract categorical regularities during statistical learning. *Dev. Sci.* **2021**, *24*, e13072. [CrossRef]
113. Student. The probable error of a mean. *Biometrika* **1908**, *6*, 1–25. [CrossRef]
114. Spitzer, B.; Waschke, L.; Summerfield, C. Selective overweighting of larger magnitudes during noisy numerical comparison. *Nat. Hum. Behav.* **2017**, *1*, 0145. [CrossRef] [PubMed]
115. Dehaene, S.; Dehaene-Lambertz, G.; Cohen, L. Abstract representations of numbers in the animal and human brain. *Trends Neurosci.* **1998**, *21*, 355–361. [CrossRef]
116. Eckert, J.; Call, J.; Hermes, J.; Herrmann, E.; Rakoczy, H. Intuitive statistical inferences in chimpanzees and humans follow Weber's Law. *Cognition* **2018**, *180*, 99–107. [CrossRef] [PubMed]
117. Libertus, M.E.; Brannon, E.M. Behavioral and neural basis of number sense in infancy. *Curr. Dir. Psychol. Sci.* **2009**, *18*, 346–351. [CrossRef]
118. Dehaene, S. Origins of mathematical intuitions: The case of arithmetic. *Ann. N. Y. Acad. Sci.* **2009**, *1156*, 232–259. [CrossRef] [PubMed]
119. Wickens, C.D.; Helton, W.S.; Hollands, J.G.; Banbury, S. *Engineering Psychology and Human Performance*, 5th ed.; Routledge: New York, NY, USA, 2021; ISBN 978-1-00-317761-6.
120. Koerber, S.; Osterhaus, C.; Sodian, B. Diagrams support revision of prior belief in primary-school children. *Frontline Learn. Res.* **2017**, *5*, 76–84. [CrossRef]
121. Siegler, R.S.; Liebert, R.M. Acquisition of formal scientific reasoning by 10- and 13-year-olds: Designing a factorial experiment. *Dev. Psychol.* **1975**, *11*, 401–402. [CrossRef]
122. Trickett, S.B.; Trafton, J.G.; Schunn, C.D. How do scientists respond to anomalies? Different strategies used in basic and applied science. *Top. Cogn. Sci.* **2009**, *1*, 711–729. [CrossRef] [PubMed]
123. Godau, C.; Vogelgesang, T.; Gaschler, R. Perception of bar graphs—A biased impression? *Comput. Hum. Behav.* **2016**, *59*, 67–73. [CrossRef]
124. Pinker, S. A theory of graph comprehension. In *Artificial Intelligence and the Future of Testing*; Freedle, R.O., Ed.; Psychology Press: New York, NY, USA, 1990; pp. 73–126, ISBN 978-0-8058-0117-0.
125. Zacks, J.M.; Tversky, B. Event structure in perception and conception. *Psychol. Bull.* **2001**, *127*, 3–21. [CrossRef]
126. Köksal, Ö.; Sodian, B.; Legare, C.H. Young children's metacognitive awareness of confounded evidence. *J. Exp. Child Psychol.* **2021**, *205*, 105080. [CrossRef]
127. Tschirgi, J.E. Sensible reasoning: A hypothesis about hypotheses. *Child Dev.* **1980**, *51*, 1–10. [CrossRef]
128. Samarapungavan, A.; Mantzicopoulos, P.; Patrick, H. Learning science through inquiry in kindergarten. *Sci. Educ.* **2008**, *92*, 868–908. [CrossRef]
129. Inhelder, B.; Piaget, J. *The Growth of Logical Thinking from Childhood to Adolescence: An Essay on the Construction of Formal Operational Structures*; Basic Books: New York, NY, USA, 1958; ISBN 978-0-415-21002-7.
130. Crowell, A.; Kuhn, D. Developing dialogic argumentation skills: A 3-year intervention study. *J. Cogn. Dev.* **2014**, *15*, 363–381. [CrossRef]
131. Kuhn, D.; David, D., Jr. Metacognition: A bridge between cognitive psychology and educational practice. *Theory Pract.* **2004**, *43*, 268–273. [CrossRef]
132. Garcia-Mila, M.; Andersen, C. Cognitive foundations of learning argumentation. In *Argumentation in Science Education: Perspectives from Classroom-Based Research*; Erduran, S., Jiménez-Aleixandre, M.P., Eds.; Science & Technology Education Library; Springer: Dordrecht, The Netherlands, 2007; pp. 29–45, ISBN 978-1-4020-6670-2.
133. Hermansen, T.K.; Ronfard, S.; Harris, P.L.; Zambrana, I.M. Preschool children rarely seek empirical data that could help them complete a task when observation and testimony conflict. *Child Dev.* **2021**, *92*, 2546–2562. [CrossRef]
134. Jirout, J.; Klahr, D. Children's scientific curiosity: In search of an operational definition of an elusive concept. *Dev. Rev.* **2012**, *32*, 125–160. [CrossRef]
135. Jirout, J.; Zimmerman, C. Development of science process skills in the early childhood years. In *Research in Early Childhood Science Education*; Cabe Trundle, K., Saçkes, M., Eds.; Springer: Dordrecht, The Netherlands, 2015; pp. 143–165, ISBN 978-94-017-9504-3.
136. Klahr, D.; Chen, Z.; Toth, E.E. Cognitive development and science education: Ships that pass in the night or beacons of mutual illumination? In *Cognition and Instruction: Twenty-Five Years of Progress*; Lawrence Erlbaum Associates Publishers: Mahwah, NJ, USA, 2001; pp. 75–119, ISBN 978-0-8058-3823-7.
137. McComas, W.F.; Olson, J.K. The Nature of Science in international science education standards documents. In *The Nature of Science in Science Education: Rationales and Strategies*; McComas, W.F., Ed.; Science & Technology Education Library; Springer: Dordrecht, The Netherlands, 2002; pp. 41–52, ISBN 978-0-306-47215-2.

138. Osborne, J.; Collins, S.; Ratcliffe, M.; Millar, R.; Duschl, R. What “ideas-about-science” should be taught in school science? A Delphi study of the expert community. *J. Res. Sci. Teach.* **2003**, *40*, 692–720. [CrossRef]
139. Priemer, B.; Hellwig, J. Learning about measurement uncertainties in secondary education: A model of the subject matter. *Int. J. Sci. Math. Educ.* **2018**, *16*, 45–68. [CrossRef]
140. Duschl, R.A.; Grandy, R. Two views about explicitly teaching Nature of Science. *Sci. Educ.* **2013**, *22*, 2109–2139. [CrossRef]
141. Fay, A.L.; Klahr, D. Knowing about guessing and guessing about knowing: Preschoolers’ understanding of indeterminacy. *Child Dev.* **1996**, *67*, 689–716. [CrossRef]
142. Metz, K.E. Emergent understanding and attribution of randomness: Comparative analysis of the reasoning of primary grade children and undergraduates. *Cogn. Instr.* **1998**, *16*, 265–285. [CrossRef]
143. Denison, S.; Xu, F. Infant statisticians: The origins of reasoning under uncertainty. *Perspect. Psychol. Sci.* **2019**, *14*, 499–509. [CrossRef]
144. O’Grady, S.; Xu, F. The development of nonsymbolic probability judgments in children. *Child Dev.* **2020**, *91*, 784–798. [CrossRef] [PubMed]
145. Silberzahn, R.; Uhlmann, E.L.; Martin, D.P.; Anselmi, P.; Aust, F.; Awtrey, E.; Bahník, Š.; Bai, F.; Bannard, C.; Bonnier, E.; et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Adv. Methods Pract. Psychol. Sci.* **2018**, *1*, 337–356. [CrossRef]
146. Gigerenzer, G. Why heuristics work. *Perspect. Psychol. Sci.* **2008**, *3*, 20–29. [CrossRef]
147. DeMiguel, V.; Garlappi, L.; Nogales, F.J.; Uppal, R. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Manag. Sci.* **2009**, *55*, 798–812. [CrossRef]
148. Gigerenzer, G.; Gaissmaier, W. Heuristic decision making. *Ann. Rev. Psychol.* **2011**, *62*, 451–482. [CrossRef] [PubMed]
149. Čavojová, V.; Šrol, J.; Jurkovič, M. Why should we try to think like scientists? Scientific reasoning and susceptibility to epistemically suspect beliefs and cognitive biases. *Appl. Cogn. Psychol.* **2020**, *34*, 85–95. [CrossRef]
150. Beilock, S.L.; DeCaro, M.S. From poor performance to success under stress: Working memory, strategy selection, and mathematical problem solving under pressure. *J. Exp. Psychol. Learn. Mem. Cogn.* **2007**, *33*, 983–998. [CrossRef] [PubMed]
151. Raab, M.; Gigerenzer, G. The power of simplicity: A fast-and-frugal heuristics approach to performance science. *Front. Psychol.* **2015**, *6*, 1672. [CrossRef]
152. Ahn, W.; Kalish, C.W.; Medin, D.L.; Gelman, S.A. The role of covariation versus mechanism information in causal attribution. *Cognition* **1995**, *54*, 299–352. [CrossRef]
153. Amsel, E.; Brock, S. The development of evidence evaluation skills. *Cogn. Dev.* **1996**, *11*, 523–550. [CrossRef]
154. Blanco, N.J.; Sloutsky, V.M. Adaptive flexibility in category learning? Young children exhibit smaller costs of selective attention than adults. *Dev. Psychol.* **2019**, *55*, 2060–2076. [CrossRef] [PubMed]
155. Nokes, T.J.; Schunn, C.D.; Chi, M. Problem solving and human expertise. In *International Encyclopedia of Education*; Elsevier Ltd.: Amsterdam, The Netherlands, 2010; pp. 265–272, ISBN 978-0-08-044894-7.
156. Konold, C.; Robinson, A.; Khalil, K.; Pollatsek, A.; Well, A.; Wing, R.; Mayr, S. Students’ use of modal clumps to summarize data. In Proceedings of the Sixth International Conference on Teaching Statistics, Cape Town, South Africa, 7–12 July 2002; p. 6.
157. Saffran, A.; Barchfeld, P.; Alibali, M.W.; Reiss, K.; Sodian, B. Children’s interpretations of covariation data: Explanations reveal understanding of relevant comparisons. *Learn. Instr.* **2019**, *59*, 13–20. [CrossRef]
158. Saffran, A.; Barchfeld, P.; Sodian, B.; Alibali, M.W. Children’s and adults’ interpretation of covariation data: Does symmetry of variables matter? *Dev. Psychol.* **2016**, *52*, 1530–1544. [CrossRef]
159. McNeil, N.M.; Alibali, M.W. Knowledge change as a function of mathematics experience: All contexts are not created equal. *J. Cogn. Dev.* **2005**, *6*, 285–306. [CrossRef]
160. McNeil, N.M.; Grandau, L.; Knuth, E.J.; Alibali, M.W.; Stephens, A.C.; Hattikudur, S.; Krill, D.E. Middle-school students’ understanding of the equal sign: The books they read can’t help. *Cogn. Instr.* **2006**, *24*, 367–385. [CrossRef]
161. Morris, B.J.; Masnick, A.M.; Baker, K.; Junglen, A. An analysis of data activities and instructional supports in middle school science textbooks. *Int. J. Sci. Educ.* **2015**, *37*, 2708–2720. [CrossRef]
162. Siegler, R.S.; Im, S.; Schiller, L.K.; Tian, J.; Braithwaite, D.W. The sleep of reason produces monsters: How and when biased input shapes mathematics learning. *Ann. Rev. Dev. Psychol.* **2020**, *2*, 413–435. [CrossRef]
163. Shtulman, A.; Walker, C. Developing an understanding of science. *Ann. Rev. Dev. Psychol.* **2020**, *2*, 111–132. [CrossRef]

Article

Analysis of Data-Based Scientific Reasoning from a Product-Based and a Process-Based Perspective

Sabine Meister * and Annette Upmeier zu Belzen * 

Biology Education, Humboldt-Universität zu Berlin, D-10099 Berlin, Germany

* Correspondence: sabine.meister@hu-berlin.de (S.M.); annette.upmeier@biologie.hu-berlin.de (A.U.z.B.)

Abstract: In this study, we investigated participants' reactions to supportive and anomalous data in the context of population dynamics. Based on previous findings on conceptions about ecosystems and responses to anomalous data, we assumed a tendency to confirm the initial prediction after dealing with contradicting data. Our aim was to integrate a product-based analysis, operationalized as prediction group changes with process-based analyses of individual data-based scientific reasoning processes to gain a deeper insight into the ongoing cognitive processes. Based on a theoretical framework describing a data-based scientific reasoning process, we developed an instrument assessing initial and subsequent predictions, confidence change toward these predictions, and the subprocesses data appraisal, data explanation, and data interpretation. We analyzed the data of twenty pre-service biology teachers applying a mixed-methods approach. Our results show that participants tend to maintain their initial prediction fully or change to predictions associated with a mix of different conceptions. Maintenance was observed even if most participants were able to use sophisticated conceptual knowledge during their processes of data-based scientific reasoning. Furthermore, our findings implicate the role of confidence changes and the influences of test wiseness.

Citation: Meister, S.; Upmeier zu Belzen, A. Analysis of Data-Based Scientific Reasoning from a Product-Based and a Process-Based Perspective. *Educ. Sci.* **2021**, *11*, 639. <https://doi.org/10.3390/educsci11100639>

Academic Editors: Moritz Krell, Andreas Vorholzer and Andreas Nehring

Received: 15 August 2021
Accepted: 8 October 2021
Published: 14 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: scientific reasoning; anomalous data; balance of nature metaphor

1. Introduction

Developing, understanding, and critically questioning knowledge and processes of deriving knowledge in science are key aspects of scientific reasoning [1,2]. The ability to engage in scientific reasoning requires a set of competences and knowledge entities that vary depending on the kind of problem to be solved [1,3]. Kind and Osborne [3] describe styles of reasoning that are distinguishable based on typical entities of conceptual, procedural, and epistemic knowledge. Conceptual knowledge focusses on the scientific objects of the problem's context, procedural knowledge focusses on entities that address methods and tools used for generating information like empirical data, and epistemic knowledge focusses on entities used to justify scientific conclusions on a meta-level [1,3–5].

Most processes of scientific reasoning rely on empirical data derived from methods like experimentation, observation, or modeling [3]. Therefore, reasoning based on data is central in scientific practices and defined as one epistemic activity in scientific reasoning [6,7]. Especially data that are not in line with prior knowledge, so-called anomalous or contradicting data [8], are a driving force for engaging in scientific reasoning. Reasoning processes initiated by anomalous data address conceptual knowledge regarding conceptual development, procedural knowledge regarding questions of methodology, and epistemic knowledge regarding questions of credibility and limits of data-based knowledge acquisition (e.g., [8–10]).

Most studies investigating reasoning in the light of anomalous data focus the analysis on participants' explanations for their reaction to the data (e.g., [8,9]), not including an analysis of the reasoning process itself. The reaction to the data can be regarded as the product from a previous reasoning process (e.g., [10,11]). Hence, studies that only focus

on the reaction (e.g., change of initial theory) analyze responses to anomalous data from a product-based view. In contrast, studies that analyze the reasoning process leading to these reactions are considered to apply a process-based view (e.g., [10,11]).

Studies that investigated responses to anomalous data from a process-based view mostly used data that were self-generated by the participants in laboratory settings [10,12]. However, reasoning processes with first-hand or second-hand data differ regarding used entities of conceptual and procedural knowledge [13].

The aim of this paper is to provide an integrational perspective from a product-based and a process-based analysis of reasoning processes with second-hand anomalous and supportive data. Therefore, reasoning processes are described by applying a general model of information processing [14] resulting in the model of data-based scientific reasoning.

The results might help to gain a deeper insight into processes that occur when reasoning with anomalous and supportive data as well as the relation to the use of conceptual, procedural, and epistemic knowledge. Further research might tie in these findings, leading to instructional recommendations for data-based scientific reasoning when used in teaching and learning.

2. Theoretical Background

2.1. Data-Based Scientific Reasoning

Chinn and Brewer [15] highlight the initiating effects of anomalous data for the development of scientific knowledge by reviewing historical examples in which anomalous data played a crucial role in the investigations of scientists leading to discussions that initiated a critical reflection on initial interpretations and theories. “Anomalous evidence are data which would not be predicted by, and are inconsistent with, a person’s mental model” [8], hence they can be described as initiators of cognitive conflicts that induce conceptual development and reasoning processes [16]. However, previous studies on anomalous data show that data contradicting initial expectations are discounted in different ways [8,15,17,18]. Such responses to anomalous data rely on a variety of justifications [8,9] based on different aspects of conceptual, procedural, or epistemic knowledge [3]. Furthermore, evidence exists that shows the importance of the perception and recognition of the anomalous data for subsequent reasoning processes [10,13,19]. More recently, a study on anomalous data provided evidence that the degree of anomaly relates to the likelihood of theory change [20]. In this study, the researchers could show that an increase of shown anomalous data increases the recognition of the anomaly and subsequently decreases participants’ confidence in the initial theory. This change in confidence was furthermore connected to a tendency to change their initial theory based on the new information provided by the anomalous data presented [20].

Responses to anomalous data are often conceptualized as part of interpretational processes during data-based scientific reasoning [21]. Previous studies show a tendency for a product-based view on responses to anomalous data and a concentration on a rather meta-level appraisal of this kind of data, asking for the believability and relevance [8,22] instead of asking for the coordination between anomalous data and initial knowledge. However, knowledge about the processes involved in different situations of scientific reasoning can lead to deeper insights into the structure of reasoning processes and enhances the knowledge about scientific reasoning [3,11].

From a process-based view, reasoning initiated by anomalous data can be described based on a general model of information processing [14], emphasizing the roles of data perception, data selection, data appraisal, data explanation, and data interpretation regarding initial knowledge (Figure 1).

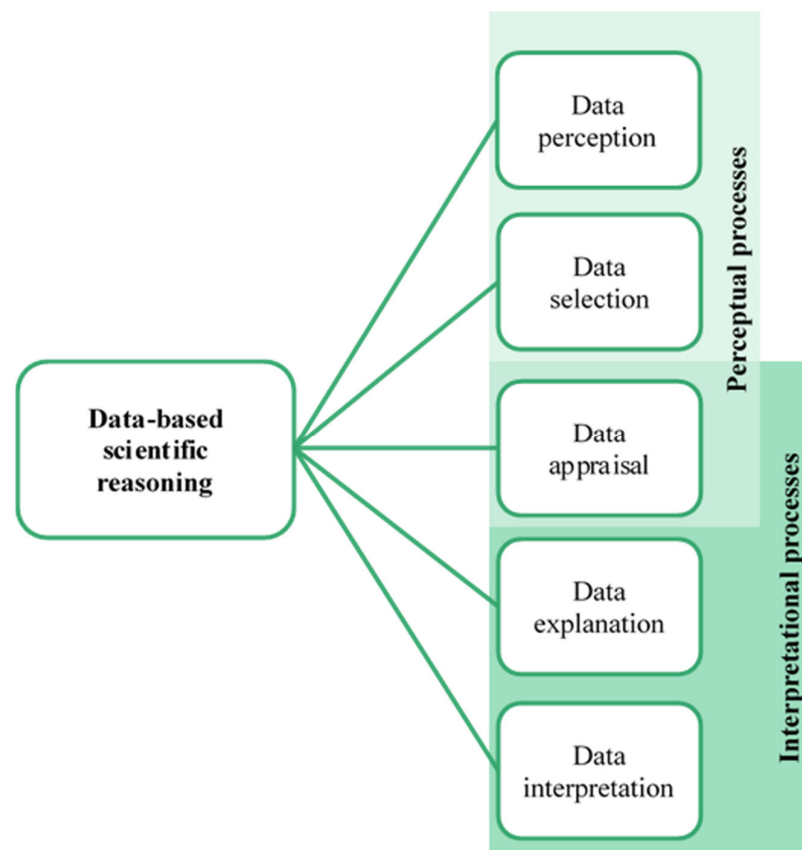


Figure 1. Theoretical model of data-based scientific reasoning (based on [8,14,19,21,23–26]).

In this process model of data-based scientific reasoning, anomalous data function as sensory stimuli that, at first, have to be perceived [10,12,19,23] before they are selected and appraised in early reasoning processes which focus on the perception of data characteristics [24,25]. Subsequently, data are interpreted within and integrated into initial knowledge entities during interpretational reasoning processes [24,27]. Interpretational processes can be distinguished into data explanation and data interpretation. Data explanation focusses on the sense-making of the data by offering alternative causes, whereas the interpretation of the data includes the coordination of the data, the alternative explanations, and the initial hypothesis to make a claim that is justified [28]. All of these sub-processes are influenced consciously or unconsciously by initially held entities of conceptual, procedural, and epistemic knowledge [3].

Research on information processing shows that a strong tendency to confirm prior conceptions can influence each step in the information processing process [29]. Therefore, we assume that responses to anomalous data, representing a specific type of scientific information, differ qualitatively in relation to the phase of information processing. Such strategies of confirmation can occur during several processes during data-based scientific reasoning, for example: perceptually ignoring contradicting data in the process of data perception, searching for flaws in contradicting data or information in the process of data appraisal, being more willing to advance vague, nonspecific causes, or finding alternative causes in the process of data explanation [18,24]. Therefore, a detailed look at the responses to anomalous data in relation to the phases of information processing provides a deeper understanding behind the cognitive processes during data-based reasoning.

2.2. Changes of Conceptual Development with Data in the Context of Population Dynamics

The acquisition of knowledge in the context of ecology is influenced by initial conceptions that are often not in line with current scientific theories [30], such as the assumption that ecosystems have a specific equilibrium state given by nature [31]. Most of these

not scientifically adequate conceptions derive from the use of the so-called Balance of Nature (BoN) metaphor [32]. Within this metaphor, ecosystems are defined as being stable, homogenous entities that regenerate to an ideal equilibrium state after disturbances. Human interactions with ecosystems are mostly seen as destructive leading to instability. According to BoN, organisms in ecosystems behave harmonically and control each other in a balanced way [32]. Conceptions on ecosystem and population dynamics that are related to BoN are prominently used in media like news, the Internet [31], and schoolbooks [33]. Therefore, it is not surprising that BoN conceptions are stable against teaching interventions [34]. The aim of teaching interventions is to initiate conceptual development by offering alternative scientifically adequate conceptions that would fit into a Flux of Nature (FoN) metaphor [31,32] and support the preference of using FoN conceptions over the BoN conception during scientific reasoning [35].

Using the example of population dynamics, the advantages, and difficulties for data-based scientific reasoning initiated by anomalous data can be shown. The development of a population in size and composition over time is a typical topic discussed in school biology and university level ecology courses [36]. However, entities of conceptual knowledge emerge from teaching interventions, but are influenced by initially held conceptions about the topic [37]. Furthermore, population dynamics are often represented by using data depicted as line graphs [38] to show, for example, the development of the population size of a species over time. Additionally, the presentation of empirical data sets is more likely to induce theory change [39]; hence, presenting anomalous data in the context of population dynamics in their typical representation as line graphs might give interesting insights for research on data-based scientific reasoning. Thus, scientific reasoning processes in this context require the use of procedural knowledge regarding handling data (e.g., knowing procedures of data generation, identifying patterns in data sets [25,26]) and interpreting graphs (diagram competence [40]). Connected to procedural knowledge, knowledge on the limits of interpreting the data are necessary for scientific reasoning, which is part of epistemic knowledge. In the case of population dynamics, represented line graphs are often connected to the use of the Lotka–Volterra equations modeling the development of populations in a prey–predator relationship hypothetically [32,41]. Therefore, epistemic knowledge associated with meta-modeling knowledge is also required during scientific reasoning in the context of population dynamics [42].

2.3. Aim and Research Questions

The aim of the following study is the identification and empirical description of reactions to anomalous and supportive data and their relation to individual processes of data-based scientific reasoning in the field of ecology. Therefore, we focused on the following research questions.

1. How does anomalous data affect the change of initial predictions regarding the scientific phenomenon of population dynamics?
2. How are changes of initial predictions about population dynamics related to a change in confidence towards the initial predictions?
3. How are reactions regarding initial predictions about population dynamics related to presented proportions of anomalous to supportive data?
4. How are reactions regarding initial predictions about population dynamics related to individual processes of data-based scientific reasoning?

3. Materials and Methods

The study is based on a mixed-methods design encompassing assessment instruments that allow the application of quantitative and qualitative analysis methods [43]. A traditional paper-and-pencil format was combined with the use of eye-tracking techniques [44]. Participants were invited to participate in the study that was conducted in a laboratory setting in the university.

3.1. Participants

In the study, twenty pre-service biology teachers (mean age = 26.25 years; SD = 5.44 years) ranging from attending first-year bachelor courses ($n_{\text{Bachelor}} = 11$) to attending master courses ($n_{\text{Master}} = 9$) participated voluntarily. The range of invited participants was chosen to enhance the variety of assessable responses to anomalous data during the process of data-based scientific reasoning due to their assumed differences in expertise regarding ecology and scientific reasoning [45].

3.2. Instrument

We developed a paper-and-pencil instrument in the context of population dynamics containing a set of tasks for assessing individual initial expectations and subsequently responding to anomalous and supportive data (Table 1). To interpret the answers given in the instrument, regarding responses to anomalous data, individual initial expectations on population dynamics were assessed by a prediction task in which participants graphed predicted outcomes of population development over a period of ten years and explained their prediction in an open-ended writing task. The prediction task was combined with a confidence rating scale for all scenarios prior to the remaining set of tasks (Table 1). Each of the following tasks is aiming to operationalize one of the sub-processes of the process model of data-based scientific reasoning (Figure 1). Perceptual processes of data-based scientific reasoning were operationalized in the paper-pencil instrument by the data selection task, which was combined with the assessment of eye-tracking data for validation purposes [44]. Interpretational processes were assessed by the data appraisal task, data explanation task, and data interpretation task (Table 1). Changes in the confidence regarding the initial predictions were assessed by a second confidence rating scale [20]).

Table 1. Overview of the used tasks and their corresponding sub-processes of the model of data-based scientific reasoning.

Sub-Process/Task	Task Content	Format of Data Assessment
Prediction	Making predictions about population development	Open-ended graphing task combined with open-ended writing task for explanation
	Rating the confidence in the made predictions	Rating scale: percentage scale from 0% (totally unconfident) to 100% (totally confident)
Data visual perception (perceptual)	Looking on the presented data sets without a further instruction.	Eye tracking experiment
Data selection (perceptual)	Selecting data sets	Multiple-choice task
Data appraisal (perceptual/interpretational)	Rating credibility, relevance, and fit of each data set	Rating scales from 1 (credible/relevant/fitting) to 5 (non-credible/irrelevant/not fitting)
Data explanation (interpretational)	Explaining each data set	Open-ended writing task
Data interpretation (interpretational)	Interpreting data sets regarding initial conceptions	Open-ended writing task
	Rating the confidence in the made predictions retrospectively	Rating scale: percentage scale from 0% (totally unconfident) to 100% (totally confident)

The contexts of the three scenarios were closely comparable with all introducing a population of an herbivorous mammal species (elk, deer, and goat) in a terrestrial ecosystem and a typical predator species. The scenarios varied regarding the proportion of anomalous and supportive data shown to induce the data-based scientific reasoning process. Anomalous and supportive data were operationalized as data sets represented as line graphs. Each of the line graphs was pre-defined to show either a population dynamic associated with typical BoN expectations (stable, slightly fluctuating population number) or typical FoN expectations (chaotic fluctuating population number, extinction; [41]). In

each scenario (elk, deer, and goat), six of these data sets were presented as a stimulus to induce the scientific reasoning process (Figure 2).

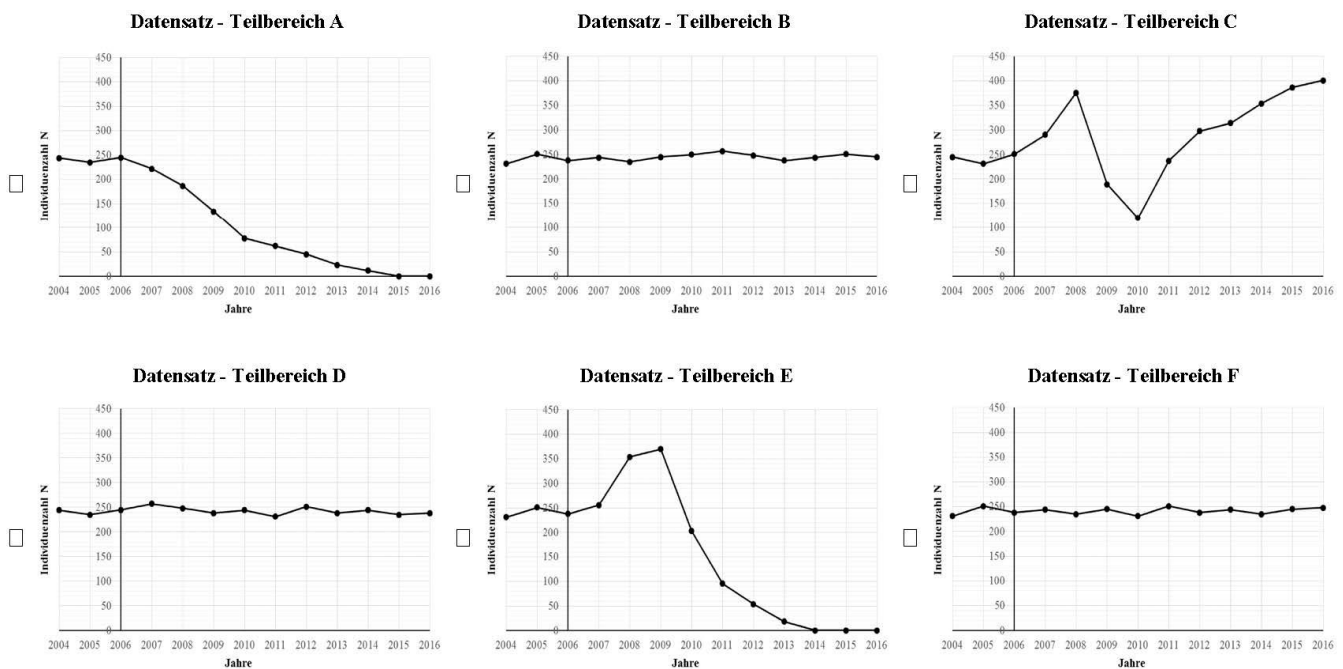


Figure 2. Example stimulus showing the six line graphs that represent different outcomes for the population development of a specific species in a defined ecosystem. Three line graphs are pre-defined as BoN-associated (B,D,F) and three line graphs are pre-defined as FoN-associated (A,C,E).

The degree of anomaly was varied by changing the ratio between FoN and BoN associated graphs from 2:4; 3:3 to 4:2 within the three scenarios [20]. Each scenario was assigned to a specific ratio between FoN and BoN-associated graphs (deer = 3 FoN:3 BoN; goat = 2 FoN:4BoN; elk = 4 FoN:2 BoN). The sequencing of the three scenarios was randomized between the participants to avoid sequencing effects. Hence, participants responded to the set of tasks three times while processing the three scenarios in different orders.

3.3. Analyses

In this study, responses to anomalous data were analyzed from a product-based and a process-based view (e.g., [10,11]). The product-based view focuses on the change of initial predictions made by the participants after reasoning with anomalous data. Therefore, the analysis is grounded strongly in the nature of the three predictions made by the participants as part of the instrument. Therefore, we coded the type of graphed prediction and associated written explanation following a qualitative content analysis approach [46]. We developed a category system that includes deductively generated categories from the main theoretical frameworks addressing conceptual, procedural, and epistemic knowledge entities that might be used when reasoning with anomalous data in the context of population dynamics [3,24–26,41]. After piloting the category system, descriptions were refined and inductively generated categories included, resulting in a final category system with 26 codes for coding the answers of all tasks included in the instrument (Table A1). The first author coded all answers from the participants. To check for the objectivity of the category system, a second coder who was no expert in this field of research re-coded 20% of the material, resulting in an intercoder agreement of $\kappa = 0.73$, indicating a good objectivity. However, disagreements were subsequently discussed and coding descriptions in the coding manual adjusted. To group the given answers of the prediction task into prediction groups, we used an epistemic network analysis (ENA [47]), using an open-source online tool that quantifies, visualizes, and models networks between qualitative entities

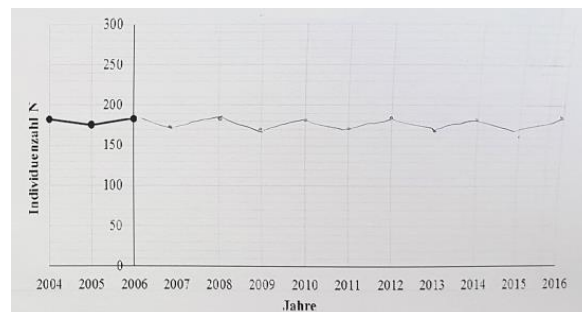
of processes such as discussions. This tool allows unraveling relations between cognitive knowledge entities and is based on theoretical frameworks for learning analytics [47]. ENA represents relations between objects in dynamic networks in which also the strength of each relation is considered [47]. Objects are represented as knot points and relations as lines between these knots varying in their thickness to indicate the strength of the relation. Objects are defined as the coded categories that indicate the use of conceptual (e.g., mentioning theories of prey–predator relationships), procedural (e.g., using statistics), and epistemic (e.g., credibility of data) knowledge entities (Appendix Table A1). Hence, each answer from the prediction task for the three scenarios per participant resulted in an individual network ($N = 60$), with the coding categories as objects and their co-occurrences as relations. All networks are located in a two-dimensional coordinate system; hence, all objects have the same position in the coordinate system independent from the individual network making different networks comparable [47]. Hence, similar networks are located closer to one another than networks that differ in their included objects and relations. To group the networks, we first distinguished the answers based on the type of graphed prediction into BoN-associated (Figure 3a,b), FoN-associated (Figure 3c), or FoN/BoN, when participants graphed two different predictions that were associated with both BoN and FoN [41]. These three groups were labeled as superior prediction groups indicating the superficial tendency of the conception behind the made prediction.

Within these superior prediction groups, similar individual networks were grouped, based on the co-occurrence of knowledge entities used for explaining the graphed predictions (represented in the ENA model as relations between objects) and labeled as explicit prediction groups. Based on this grouping, summary statistics that are included to ENA allow an aggregation of all networks in a group into a mean network. Hence, a mean network represents the average combination of objects and their relations for this group [47]. In this study, mean networks of an explicit prediction group showed typical combinations of used knowledge entities for explaining the made prediction regarding population development. Furthermore, ENA offers the calculation of t -tests (e.g., Mann–Whitney test) to check for a statistically significant difference between the mean networks of different groups [47].

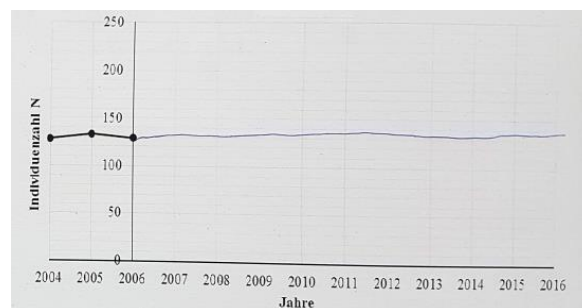
Based on the found prediction groups, we observed if participants changed the prediction group for the second and third scenario in the instrument after reasoning with anomalous and supportive data regarding their initial prediction (Figure 4; prediction group change). Furthermore, changes of confidence in the initial prediction (Figure 4; confidence change) and the relation to the presented proportion of anomalous to supportive data were taken into consideration as factors that might influence the responses to anomalous data.

Subsequently to this product-based view of analysis, we analyzed the data-based reasoning processes that occurred between the prediction group changes and confidence changes (Figure 1 DbR processes). For this process-based analysis (e.g., [10,11], answers to the data appraisal task, data explanation task, and data interpretation task were analyzed for the first and second scenario of each participant. We excluded the third scenario in this analysis since we did not assess a further prediction change after the reasoning process during the third scenario due to the test construction. The answers of the rating scales in the data appraisal task were subsumed into five groups. If participants rated the credibility and the relevance of the perceived anomalous data as low (1 or 2 on the rating scale) they were assigned to *skeptical general*. When participants rated the perceived anomalous data as only low on the credibility scale, they were assigned to *skeptical credibility*; in the case of the relevance scale this led to *skeptical relevance*. Participants who rated both scales in the middle (3 on the rating scale), were assigned to *undecided*, and participants who rated high on both scales (4 and 5 on the rating scale) were assigned to *not skeptical*. After coding the answers to the open-ended questions from the data explanation task and data interpretation task, we compared the used conceptual knowledge entities with the ones the participants used for their prediction in each scenario. Based on this comparison, two groups were defined as *new conceptual knowledge* and *initial conceptual knowledge*. *New conceptual knowledge*

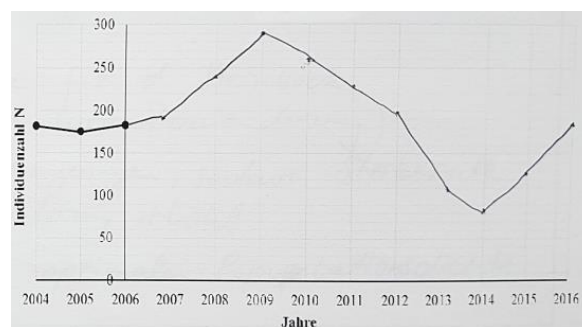
encompasses cases in which participants used new conceptual knowledge entities in addition to the initial conceptual knowledge entities, for example, when a participant used theories of prey–predator relationships for their prediction only but explained or interpreted the data by considered environmental factors like natural resources. *Initial conceptual knowledge* encompasses cases in which participants only used initial conceptual knowledge entities, for example, when the previous mentioned participant used theories of prey–predator relationships during data explanation and interpretation as the single explanation option. If participants additionally used procedural or epistemic knowledge entities for explaining and interpreting data, they were assigned to the sub-groups *plus procedural* or *epistemic knowledge*. Participants that answered without using conceptual, procedural, or epistemic knowledge to explain or interpret data were assigned to *no explanation*. Based on this grouping, participants’ data-based scientific reasoning processes were assigned into a dimensional matrix with data appraisal on one dimension and data explanation/interpretation on the other dimension.



A



B



C

Figure 3. (a,b) Examples of graphed predictions for the population development of a specific species in a defined ecosystem that were assigned into BoN-associated. (c) Example of a graphed prediction for the population development of a specific species in a defined ecosystem that was assigned into FoN-associated.

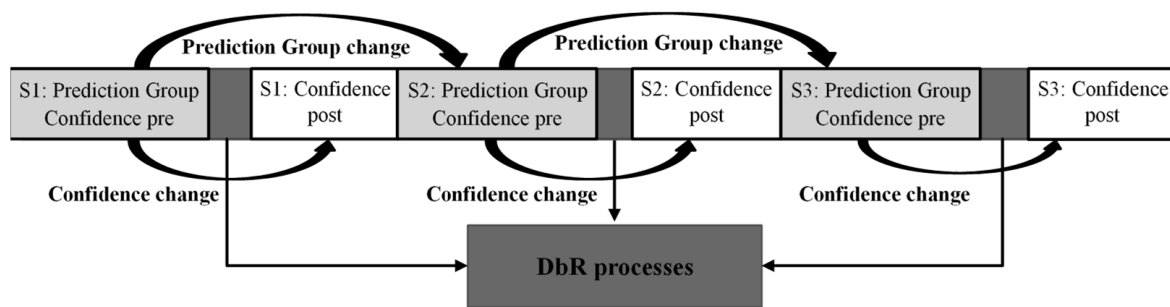


Figure 4. Schematic representation of the analysis processes for this study.

4. Results

Each of the participants ($N = 20$) answered the prediction and data-based scientific reasoning tasks (Table 1) for the three scenarios leading to a total amount of 60 answers for each task. For the open-ended writing tasks that were coded by a qualitative content analysis, a total of $N = 868$ codes were assigned, ranging from 19 to 59 codes between participants.

First, the results regarding the prediction groups found by ENA are presented. All individual networks for the answers of the prediction tasks in the three scenarios per participants ($N = 60$) were modeled into a dynamics network by ENA as shown in Figure 5.

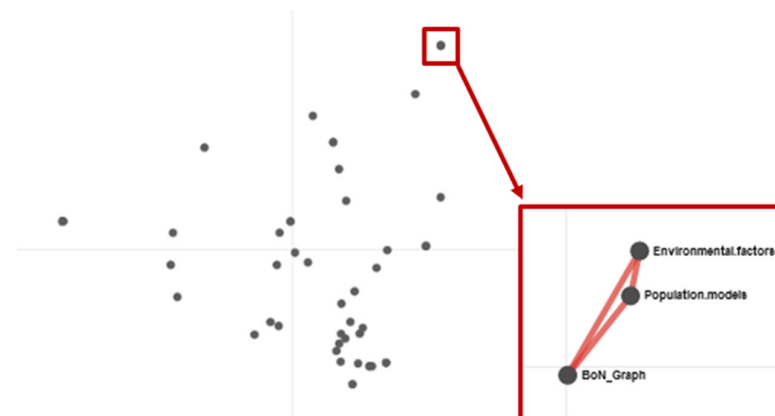


Figure 5. Individual networks for all predictions made by the participants in a two-dimensional system modeled with ENA.

From these individual networks presented as dots, seven explicit prediction groups were defined (Table 2). However, in ten individual networks that represent answers to the prediction task in the second and third scenario, the main explanation for the made prediction was test wiseness. Test wiseness is operationalized as identifying participants' statements that present experiences from the previous tasks of the test instrument as the main reasons for the task performance under consideration instead of answering the task based on conceptual, epistemic, or procedural knowledge. Test wiseness is often used to improve test performance [48]. For example: "A stable graph was shown in the previous scenario. I want to cover every option".

The Mann–Whitney test showed that explicit prediction groups within their superior group were statistically different at the $\alpha = 0.05$ level in at least one dimension of the coordinate system, except for *divergent prey–predator relation conceptions* and *mixed conceptions and human disturbance* in the FoN/BoN group (Table 2). Based on the theoretical background, both groups represent different aspects of conceptions associated with the BoN metaphor [30,32]; hence, we maintained both explicit prediction groups.

Table 2. Descriptions and absolute frequencies per scenario (N 1st; N 2nd; N 3rd) of superior prediction groups and explicit prediction groups found with ENA.

Superior Prediction Groups	Explicit Prediction Groups	Mean Network Model	Description	N 1st	N 2nd	N 3rd
	Harmonic prey–predator relation (PPR) conception		Participants assigned to this group graphed BoN predictions and explained their predictions with their content knowledge about population models that they connected with conceptions about stability and harmonic prey–predator relationships.	4	3	1
BoN	Stability conception		Participants assigned to this group graphed BoN predictions and explained their predictions with a general stability conception.	7	3	4
	Content knowledge		Participants assigned to this group graphed BoN predictions and explained their predictions with biological content knowledge. They mentioned population models and environmental factors, without connecting these with stability conceptions.	3	2	1
FoN/BoN	Mixed conceptions and content knowledge		Participants assigned to this group graphed BoN predictions and FoN predictions. They explained their predictions with biological content knowledge. They connected their knowledge with divergent conceptions addressing both FoN (natural causes, inharmonic PPR) and BoN (stability, harmonic PPR).	3	4	2

Table 2. Cont.

Superior Prediction Groups	Explicit Prediction Groups	Mean Network Model	Description	N 1st	N 2nd	N 3rd
	Divergent prey–predator relation conceptions		Participants assigned to this group graphed BoN predictions and FoN predictions. They explained their predictions with divergent conceptions about prey–predator relationships addressing both FoN and BoN.	0	1	2
	Mixed conceptions and human disturbance		Participants assigned to this group graphed FoN predictions. They explained their predictions with biological content knowledge and FoN related conceptions. They also mentioned human disturbance when explaining their predictions.	1	3	1
FoN	FoN conceptions and content knowledge		Participants assigned to this group graphed FoN predictions. They explained their predictions with biological content knowledge, mostly mentioning population models. They connected their knowledge with FoN-related conceptions.	2	0	3

Most predictions given by the participants indicate a tendency towards BoN conceptions ($n = 28$; 46.7%) or a mix of BoN and FoN conceptions ($n = 17$; 28.3%). Therefore, BoN-associated data sets presented in the instrument are assumed to be perceived as supportive, while FoN-associated data sets are assumed to be perceived as anomalous data. This assumption is supported by the decrease of frequencies for BoN prediction groups and an increase of FoN/BoN prediction groups after the first scenario (Table 2).

4.1. Prediction Group Changes

Based on the assignment of participants' answers given to the prediction task to the prediction groups for each scenario, the changes of prediction groups between scenarios were analyzed. Prediction group changes were expected between the scenarios as a reaction to reasoning with anomalous and supportive data regarding the initial prediction made in the previous scenario. Table 3 shows how many participants maintained or changed their superior prediction group from the first to second and second to third scenario.

Table 3. Absolute frequencies of superior prediction group changes between the first and second scenario and the second and third scenario.

To From	BoN	FoN/BoN	FoN
BoN	$n_{1st-2nd} = 8$ (* = 1) $n_{2nd-3rd} = 6$ (* = 1)	$n_{1st-2nd} = 6$ (* = 3) $n_{2nd-3rd} = 1$	$n_{1st-2nd} = 0$ $n_{2nd-3rd} = 2$
FoN/BoN	$n_{1st-2nd} = 0$ $n_{2nd-3rd} = 3$ (* = 2)	$n_{1st-2nd} = 4$ $n_{2nd-3rd} = 7$ (* = 3)	$n_{1st-2nd} = 0$ $n_{2nd-3rd} = 1$
FoN	$n_{1st-2nd} = 1$ $n_{2nd-3rd} = 0$	$n_{1st-2nd} = 1$ $n_{2nd-3rd} = 0$	$n_{1st-2nd} = 0$ $n_{2nd-3rd} = 0$

* Frequencies of cases in which test wiseness was included into the explanations for a made prediction.

In most possible changes ($n = 40$) the initial prediction groups were maintained, especially when BoN conceptions ($n = 14$; 35%) or a mix of FoN and BoN conceptions ($n = 11$; 27.5%) were used initially in the prediction task. Changes of prediction groups between the scenarios occurred fifteen times (37.5%). Most of the changes occurred from prediction groups associated with BoN conceptions to prediction groups associated to a mix of FoN and BoN conceptions ($n = 7$; 17.5%). In four cases (10%), a change from an FoN or mixed-associated prediction to a more BoN-associated prediction occurred. In particular, changes to and the maintenance of an FoN/BoN prediction group were related to the effect of test wiseness. When participants maintained the superior prediction group, they also maintained their explicit prediction group with one case as an exception.

4.2. Reactions to Anomalous Data

For each scenario, the participants rated their confidence in their prediction before and after dealing with anomalous and supportive data sets on a percentage scale. The difference between the two ratings represents the confidence change. Based on the found differences, five options of confidence change were identified: *steady confidence* when confidence remained above 50% on the rating scale, *steady unconfidence* when confidence remained under 50% on the rating scale, *confidence in abeyance* when confidence remained on 50% on the rating scale, *increase to confidence* when confidence changed from under 50% to above 50% on the rating scale, and *decrease to unconfidence* when confidence changed from above 50% to under 50% on the rating scale. Table 4 shows the frequencies of each option across the three scenarios to which the participants gave answers.

The data-based scientific reasoning process with anomalous and supportive data sets in the first scenario led to a wide range of responses regarding the confidence in the initial prediction. While some participants maintained their initial rating of confidence, either as confident or as unconfident, six participants increased their confidence in their prediction after dealing with the data. Furthermore, three participants decreased their confidence, and four participants were undecided about their confidence. In contrast, the frequencies of the confidence change options for the second and third scenarios show a tendency to

maintain the rated confidence, either as confident or as unconfident, after dealing with the shown data sets representing population dynamics.

Table 4. Absolute frequencies of options for confidence change which occurred within the first, second, and third scenarios.

Confidence Change Options	N (1st Scenario)	N (2nd Scenario)	N (3rd Scenario)
Steady confidence	4	8	9
Steady unconfidence	4	5	4
Confidence in abeyance	4	4 (+1)	5
Increase to confidence	5 (+1) ¹	1	1
Decrease to unconfidence	3	2	1 (+1)

¹ One participant made two different predictions and rated them separately.

To check relations between confidence change and prediction group change, the presented frequencies shown in Table 3; Table 4 were integrated. Data from Table 4 were limited to the columns for the first and second scenarios because we assessed no further change of the prediction group after participants answered the instrument for the third scenario. Based on this data integration, we defined six possible reactions after dealing with the shown anomalous and supportive data sets (Table 5).

Table 5. Absolute frequencies and percentages of reactions to anomalous data shown by the participants.

Reactions	N
Confident confirmation	14 (* = 3; 35%)
Undecided confirmation	4 (10%)
Unconfident confirmation	7 (* = 2; 17.5%)
Confident modification	4 (* = 1; 10%)
Undecided modification	4 (* = 1; 10%)
Unconfident modification	7 (* = 3; 17.5%)

* Frequencies of cases in which test wiseness was included into the explanations for a made prediction.

Mostly, participants that maintained their prediction group were confident in their prediction after data-based scientific reasoning ($n = 14$; 35%). Still, twenty percent of participants maintained their prediction group even if they stated that they are unconfident about their prediction. If participants changed the prediction group by modifying their prediction between the first and second scenario or second and third scenario, they mostly stated to be unconfident towards their initial prediction ($n = 7$; 17.5%).

4.3. Relation to the Proportion between Anomalous Data and Supportive Data

All participants gave predictions for each of the three scenarios that differ in the proportion between presented BoN and FoN-associated data sets; hence, the proportion of perceived supportive and anomalous data varies. The three scenarios were randomly sequenced between the participants. Table 6 shows the frequencies of reactions to the data in relation to the different proportions between supportive and anomalous data also labeled as the anomalous data ratio.

For both types of reactions to the data, confirmation or modification of the initial prediction, the differences between the frequencies per anomalous data ratio are rather ambiguous showing no statistical difference. However, for confirmation, a tendency of an increasing confidence when confronted with a higher or equal proportion of FoN-associated data sets to BoN-associated data sets can be found.

Table 6. Absolute frequencies and percentages of reactions to anomalous and supportive data shown by the participants in relation to the anomalous data ratio within the three scenarios.

Reactions	Anomalous Data Ratio (BoN:FoN)		
	2:4	3:3	4:2
Confident confirmation	<i>n</i> = 5 (* = 1)	<i>n</i> = 6 (* = 2)	<i>n</i> = 3
Undecided confirmation	<i>n</i> = 2	<i>n</i> = 0	<i>n</i> = 2
Unconfident confirmation	<i>n</i> = 2	<i>n</i> = 1	<i>n</i> = 4 (* = 2)
Confirmation (N = 25)	<i>n</i> = 9 (* = 1; 36%)	<i>n</i> = 7 (* = 2; 28%)	<i>n</i> = 9 (* = 2; 36%)
Confident modification	<i>n</i> = 2 (* = 1)	<i>n</i> = 1	<i>n</i> = 1
Undecided modification	<i>n</i> = 0	<i>n</i> = 3 (* = 1)	<i>n</i> = 1
Unconfident modification	<i>n</i> = 3 (* = 2)	<i>n</i> = 2 (* = 1)	<i>n</i> = 2
Modification (N = 15)	<i>n</i> = 5 (* = 3; 33.3%)	<i>n</i> = 6 (* = 2; 40%)	<i>n</i> = 4 (26.7%)

* Frequencies of cases in which test wiseness was included into the explanations for a made prediction.

4.4. Role of Data-Based Reasoning Process

In Table 7, participants' data-based scientific reasoning processes for the first and second scenario are represented as cells in a two-dimensional system with their assignment to the data appraisal groups in the one dimension and the assignment to the explanation/interpretation groups in the other dimension.

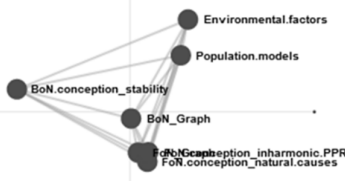
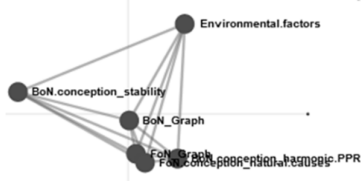
Table 7. Assignment of participants' data-based scientific reasoning processes into the two dimensions data appraisal and data explanation/interpretation based on their answers for the first and second scenario. Participants' reactions regarding their initial prediction are highlighted with italic letters when assigned to *confirmation* (*n* = 25; * = 5) and bold letters when assigned to modification (*n* = 15; * = 5).

	New Conceptual Knowledge		Initial Conceptual Knowledge		No Explanation
	Only	Plus Procedural and/or Epistemic Knowledge	Only	Plus Procedural and/or Epistemic Knowledge	
Skeptical general	Finn_1st <i>Sam_2nd</i>				
Skeptical credibility		<i>Alex_1st</i> <i>Andrea_2nd</i>			
Skeptical relevance		Andy_2nd Jamie_1st Quinn_2nd Bente_1st *			
Undecided	<i>Andrea_1st</i> <i>Bente_2nd</i> * Jona_2nd <i>Kay_2nd</i> Noah_1st Noah_2nd	<i>Chris_1st</i> <i>Chris_2nd</i> <i>Finn_2nd</i> <i>Kim_1st</i> <i>Luca_1st</i> <i>Luca_2nd</i> <i>Quinn_1st</i> <i>Sam_1st</i> <i>Alex_2nd</i> * <i>Andy_1st</i> <i>Charlie_1st</i> *	Nicola_2nd	Nicola_1st	
Not skeptical	<i>Charlie_2nd</i> <i>Jona_1st</i> <i>Kim_2nd</i> * <i>Mika_2nd</i> Toni_1st *	Jamie_2nd * <i>Kay_1st</i> Mika_1st Robin_1st * <i>Robin_2nd</i> * <i>Sascha_1st</i> <i>Sascha_2nd</i>			Toni_2nd *

* Participants' cases in which they used test wiseness as an explanation for their made prediction.



Based on this, it is shown that most of the data-based scientific reasoning processes leading to confirmation were characterized by an undecided or not skeptical appraisal of the data combined with the use of new conceptual knowledge entities in addition to the initial conceptual knowledge entities ($n = 15$; 60%). Generally, all data-based scientific reasoning processes leading to confirmation were related to the use of new conceptual knowledge entities when explaining/interpreting the data. For a deeper insight into this finding, we first looked for the assigned superior prediction groups of these cases ($n_{\text{FoN/BoN}} = 11$; $n_{\text{BoN}} = 14$). For those cases that maintained an FoN/BoN prediction group, most of the presented data sets were not anomalous, hence, there was no need for modifying the initial prediction as it was not induced by the processed data. This is illustrated by the example of Sascha (Table 8).

Table 8. Illustration of the prediction group change and data-based scientific reasoning process of Sascha in the first scenario.

Prediction Group 1st Scenario	Data Interpretation (Extract)	Prediction Group 2nd Scenario
<p>Mixed conceptions and content knowledge</p> 	<p>“During this time, factors exist that influenced the population density in a negative way (e.g., predators, disasters).”</p> <p>“Similar to prediction, only time period for regeneration of the population density was not correct.”</p> <p>“Confidence highly increased due to the similarities to the data.”</p>	<p>Mixed conceptions and content knowledge</p> 

When participants maintained their BoN prediction group, they explained or interpreted the data by using different conceptual knowledge entities but were undecided or skeptical regarding the FoN data sets (anomalous data) by tendency. The confirmation of the initial prediction was often explained by arguing with the higher ratio of supporting data sets (statistical reasoning), as exemplified by the case of Chris (Table 9).

Table 9. Illustration of the prediction group change and data-based scientific reasoning process of Chris in the second scenario.

Prediction Group 2nd Scenario	Data Interpretation (Extract)	Prediction Group 3rd Scenario
<p>Stability conception</p> 	<p>“Massive changes of environmental circumstances led to the extinction or extreme population fluctuations.”</p> <p>“In 2/3 of the areas, my prediction was the case.”</p> <p>“Without further information about environmental factors, my confidence regarding my prediction will not increase.”</p>	<p>Stability conception</p> 

Participants who modified their initial prediction showed different data-based scientific reasoning processes. For describing these cases, the direction of modification was considered ($n_{\text{FoN direction}} = 10$; $n_{\text{BoN direction}} = 5$). Almost all modifications of predictions into the FoN direction were related to data-based scientific reasoning processes in which new conceptual knowledge was used, shown by the example of Mika (Table 10).

Modifications of predictions into the BoN direction were related to data-based scientific reasoning processes with a stronger focus on procedural or epistemic knowledge like looking for statistical patterns or argumentations considering the probability of the data. This is illustrated by the example of Nicola (Table 11).

Table 10. Illustration of the prediction group change and data-based scientific reasoning process of Mika in the first scenario.

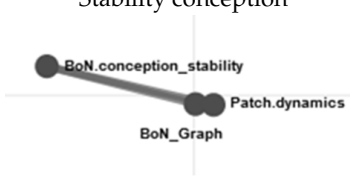
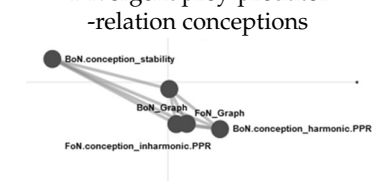
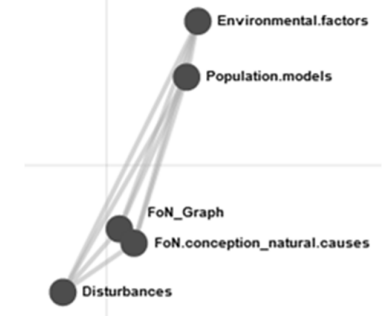
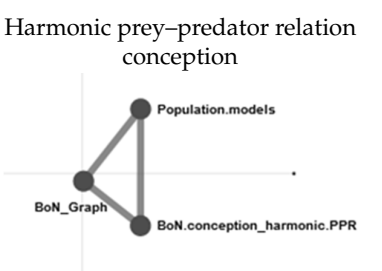
Prediction Group 1st Scenario	Data Interpretation (Extract)	Prediction Group 2nd Scenario
<p>Stability conception</p> 	<p>“4 of 6 data sets are supporting my prediction, because of a stable prey-predator relationship.”</p> <p>“2 of 6 data sets show massive fluctuations. Imbalance of prey-predator relationship could also be influenced by other factors.”</p> <p>“Unconfidence due to wrong assumptions and the fact, that population growth cannot be explained only by considering prey-predator relationships.”</p>	<p>Divergent prey-predator -relation conceptions</p> 

Table 11. Illustration of the prediction group change and data-based scientific reasoning process of Nicola in the first scenario.

Prediction Group 1st Scenario	Data Interpretation (Extract)	Prediction Group 2nd Scenario
<p>FoN conceptions and content knowledge</p> 	<p>“My prediction did not include extreme events like diseases or influences of weather, but only the development based on prey-predator-relationships.”</p> <p>“My confidence did not change, because some data represent extreme events that were not included into my prediction.”</p>	<p>Harmonic prey-predator relation conception</p> 

However, for some cases of both reaction types of confirmation and modification, test wiseness had an influence, indicating the tendency to answer the tasks of the instrument in a way that was perceived as the expected one by these participants.

5. Discussion

In this study, our aim was to investigate how participants reason with supportive and anomalous data in the context of population dynamics. In particular, we were interested in the way they confirmed or modified an initial prediction after dealing with different data sets represented as line graphs (Figure 2) by answering tasks coherent to the sub-processes of a data-based scientific reasoning process (Figure 1). For this, we integrated analyses with a product-based and a process-based view.

The first finding supports previous studies investigating conceptions about ecosystems and populations dynamics [30,34,49]. Most of the participants explained their predictions about the development of a population by using conceptions associated with the BoN metaphor (Table 2). Some participants showed a mix of BoN and the scientifically more adequate FoN metaphor-associated conceptions. Furthermore, it is shown that the frequencies of used mixed conceptions increased after the first scenario while using pure BoN conceptions decreased for making a prediction (Table 2). However, most participants maintained their initial predictions (Table 3). This finding supports the theory that conceptions are not replaced by one another, but different conceptions for a phenomenon exist parallel to each other, for example, naïve and scientifically adequate explanations for population dynamics [35]. Which conception is used in a situation depends on the characteristics of the situation itself, as this can inhibit or promote the prevalence of a specific conception [35]. In this study, participants’ conceptions associated with FoN might have been activated with the presentation of the corresponding data sets in the first scenario.

From this product-based view on the results of the study [3], we can distinguish the reactions of participants to the presented data into the confirmation or modification of the initial prediction. Both reactions are related to the confidence participants had in their initial prediction (Table 5). While confirmation is by tendency related to a high confidence in the initial prediction, modification mostly relates to a stated unconfidence in the initial prediction. These findings are consistent with the results of the study by Hemmerich and colleagues [20] in which they found that a decrease in confidence will increase the probability to change the initial theory. However, they found evidence to support the Incremental Change Hypothesis which states that the proportion of anomalous data to supportive data will influence confidence change [20]. In our study, we found by tendency opposite findings regarding the Incremental Change Hypothesis for the reaction of confirmation (Table 6). More or an equivalent proportion of FoN-associated data sets to BoN-associated data sets presented as line graphs led, by tendency, to an increased confidence in the initial prediction. However, a higher proportion of BoN-associated data sets had the opposite effect (Table 6). We assume two causes for this finding. First, predefined FoN-associated data sets, that represent a chaotic fluctuation of the population dynamic, were often interpreted in line with assumed harmonic-fluctuations and hence were perceived as supportive data for BoN predictions. This observation fits with findings of other studies which showed that some people tend to reinterpret anomalous data as fitting with their initial expectation, and hence, perceiving no anomaly at all [8]. Second, in 44% of the cases in which the initial prediction was confirmed in a subsequent scenario, the prediction was assigned into the superior prediction group FoN/BoN. Therefore, data sets that might have been perceived as anomalous were mostly limited to the data sets representing an extinction event. Furthermore, the modification of the initial prediction does not show a relation to the options of confidence change. One important reason might be that one third of the cases in which modification of the prediction occurred were based on test wiseness. Therefore, the modification shown by the participants was not motivated by processing the data in the scenario in a scientific way, but by copying the data sets as predictions to fit an expected outcome in the tasks of the subsequent scenarios. According to the finding for confidence change, this supports previous findings that show how participants' confidence is more related to the individual perception of acceptance by other people than the ability to refer to evidential considerations [50].

However, besides the effect of test wiseness during the product-based analysis, we do not know how the processing of the data sets during data-based scientific reasoning relates to the reactions regarding the initial predictions. Hence, the analyses of the tasks operationalizing the sub-processes of data-based scientific reasoning, with a focus on the interpretational processes, gave a deeper insight. Based on this, we found that the participants used mostly a combination of conceptual, procedural, and epistemic knowledge to explain and interpret data. In addition, most of them seemed undecided or not skeptical when appraising the data regarding relevance and credibility. Compared to previous studies that investigated responses to anomalous data, our study design favors responses which try to explain the data on a conceptual basis, like *reinterpretation*, *peripheral theory change*, and *theory change* in the taxonomy of responses to anomalous data [8], or *use of theoretical concepts* in the categories of justifications to hold or reject a hypothesis [9]. This is consistent with the methodological differences between our and the cited studies. First, we explicitly instructed the participants to explain each data set and interpret the data sets regarding their initial prediction. However, Chinn and Brewer [8] asked their participants to rate the believability and consistency to an initial theory of the presented data and explain their ratings. These instructions focus rather on the sub-process of data appraisal; hence, a tendency towards response types that are more on 'the data side of the [explanation] model' are expectable [24]. Second, in our study we presented second-hand data represented as line graphs. Compared to Chinn and Brewer [8] who used textual descriptions of evidence, the presentation of empirical data is typical of scientific domains. Furthermore, the representation of data as text passages [8,17], charts [51], or graphs [52]

will influence the ambiguity of the perceived anomaly. For example, Masnick and colleagues [39] gave empirical support that reasoning with numerical data initiate and support processes of conceptual change which need the activation of conceptual knowledge to formulate alternative explanations. Ludwig and colleagues [9] let participants generate data in laboratory settings or with computer simulations; therefore, they found a variety of justifications to hold or reject a hypothesis that are connected to the methodological issues of the data generation. This fits with findings of studies investigating the effect of first-hand or second-hand data on scientific reasoning. Hug and McNeill [13] concluded that first-hand data support the awareness of limitations and error in data, as well as learners' understanding of the role of data for knowledge generation in science. This is also supported by findings from other studies, investigating responses to anomalous data during experimentation and modeling activities [10,12]. Second-hand data, in turn, are perceived as authoritative by learners and support more sophisticated reasoning skills like identifying patterns, drawing conclusions, and considering content knowledge, due to being often rather complex compared to first-hand data [13]. These conclusions were supported by our findings that conceptual, procedural, and epistemic knowledge were central during participants' data-based scientific reasoning processes.

Nevertheless, sophisticated data-based scientific reasoning processes in which new conceptual knowledge is used to explain data do not lead to a change of the initial prediction *per se*. Hence, in almost all analyzed reasoning processes, new conceptual knowledge was used independent from the subsequent reaction of confirmation or modification regarding the initial prediction. Our analysis approach to integrate a product-based with a process-based view on responses to supportive and anomalous data showed that initial conceptions are strongly held and repeated even if alternative conceptions and explanations are available but are perceived as less likely due to arguments based on epistemic and procedural knowledge.

In general, scientific reasoning is proposed to rely on conceptual, procedural, and epistemic knowledge independent of the used style of reasoning that may be associated with data-based scientific reasoning or not [3]. Hence, our findings suggest that the interdependency between these forms of knowledge might be of crucial interest for future research on scientific reasoning. The role of conceptual knowledge is one aspect that has been extensively discussed lately [53]. Furthermore, a lot of research on the nature of science has been done, a construct that includes many aspects of epistemic knowledge and is related to scientific reasoning skills [54]. However, data-based scientific reasoning might be essential for most scientific reasoning styles, and it is important for all people to engage in data-based argumentation and decision making in the context of socio-scientific and controversial science issues [55].

The interpretation and generalization of the findings of this study have limitations because of methodological decisions. Due to the amount of different data sources to enable the integrational analysis, the sample size was limited. Hence, all interpretations made from the data show tendencies that need to be tested in further studies. However, with this mixed-method approach new hypotheses can be built and tested in subsequent studies. For instance, it would be interesting to observe possible causes for the tendency to maintain an initial expectation and its conceptual explanation, even if other explanations are known, but maybe seen as less likely. In addition, it might be interesting to investigate how other factors regarding data characteristics, besides the proportion between anomalous and supportive data, relate to the data-based scientific reasoning process and their outcomes. This might be moderated by a change of skepticism regarding the data. Additionally, we decided to focus the analysis of this study on the prediction group changes and corresponding data-based scientific reasoning processes, hence we presented the results of the data-based scientific reasoning processes for the first and second scenario. Furthermore, our model of data-based scientific reasoning encompasses and highlights the role of perception. This study focused on the interpretational processes during data-based scientific reasoning; however, the role of perceptual processes is still important for gaining further insights into

ongoing cognitive processes. Therefore, the analyzing of additional data assessed with eye-tracking techniques [44] will be the focus of our future research.

Author Contributions: Conceptualization, S.M. and A.U.z.B.; methodology, S.M.; formal analysis, S.M.; investigation, S.M.; writing—original draft preparation, S.M.; supervision, A.U.z.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: The respondents agreed to data use for research.

Data Availability Statement: The datasets are not publicly available.

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

Appendix A

Table A1. Category System.

Category	Subcategory	Code	Description
Type of graphed prediction		BoN Graph	The graph shows a trend that represents a stable population development. Stable is defined as linear horizontal or around a mean value fluctuating lines. The fluctuation is mostly uniform, and the amplitudes are low.
		FoN Graph	The graph shows an unstable, chaotic trend. FoN graphs include increasing, decreasing, and chaotic or with high amplitudes fluctuating graphs.
Conceptual knowledge	BoN conceptions	Stability	The general assumption of a stable development or that disturbances are not expected is stated.
		Human disturbances Harmonic prey–predator relationship (PPR) Instability	Human caused disturbances are named as reasons for instability. A harmonic regulation by prey–predator relationship is stated as a reason for stability. An unpredictable/instable development is described.
	FoN conceptions	Natural causes	Natural causes (e.g., disturbances like epidemics, fires, and invasive species; climate changes; change of environmental resource; imi- and emigration) are described as reasons for an instable development.
		Inharmonic PPR Population models	Predator caused changes that may also cause extinction are stated. Biological models like capacity limit, logarithmic population development, or prey–predator models (Lotka–Volterra) are named.
	Content knowledge	Patch dynamics	Aspects of a heterogeneous ecosystem like naturally changing resources or imi- and emigration of populations are named.
		Disturbances Biodiversity Environmental factors	The chance and importance of disturbances for development in ecosystems are named. Aspects of biodiversity (also genetics) are named. Change of biotic and/or abiotic factors are named.
Procedural knowledge	Diagram competence	Statistics	The data are statistically treated (e.g., comparison of means/data points, calculating/estimating mean values).
		CVS	Aspects of the importance to control variables are stated.
		Patterns	The identification of patterns in the data is stated.
		Represent	The data sets represented as line graphs are described superficially without explaining the shown relation.
		Syntactic	The data sets represented as line graphs are described by stating aspects of the shown relation, trend or single data points, no connection to the phenomenon/conceptual knowledge is given. Data sets are compared superficially.
	Semantic	The data sets represented as line graphs are described by stating aspects of the shown relation, trend, or single data points and a connection to the phenomenon/conceptual knowledge is given. Data sets are compared with relation to the phenomenon.	
Epistemic knowledge		Limits of models	Aspects of the limits or hypothetical nature of models are named.
		Probability Credibility	Aspects of probability and significance are named. Aspects of credibility or believability of the data are stated.
		Quality	Aspects of quality of the data are stated (e.g., reliability of measurement, replication, experimentation bias).
Others		Uncertainty Test wiseness General prior knowledge/Intuition	Aspects of uncertainty (e.g., need for more information) are stated. Experiences from previous tasks are stated as reasons for any task performance. General prior knowledge (e.g., memorizing from schoolbooks) or intuition are stated as reasons for any task performance.

References



- Fischer, F.; Kollar, I.; Ufer, S.; Sodian, B.; Hussmann, H.; Pekrun, R.; Eberle, J. Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learn. Res.* **2014**, *2*, 28–45.
- Zimmerman, C.; Croker, S. Learning science through inquiry. In *Handbook of the Psychology of Science*; Feist, G.J., Gorman, M.E., Eds.; Springer Publishing Company: New York, NY, USA, 2013; pp. 49–70.
- Kind, P.; Osborne, J. Styles of scientific reasoning: A cultural rationale for science education? *Sci. Educ.* **2017**, *101*, 8–31. [CrossRef]
- Lehrer, R.; Schauble, L. The development of scientific thinking. In *Handbook of Child Psychology and Developmental Science: Cognitive Processes*; Liben, L.S., Müller, U., Lerner, R.M., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2015; pp. 671–714.

5. Zimmerman, C. The development of scientific thinking skills in elementary and middle school. *Dev. Rev.* **2007**, *27*, 172–223. [CrossRef]
6. NGSS Lead States. *Next Generation Science Standards: For States, by States*; The National Academy Press: Washington, DC, USA, 2013.
7. Samarapungavan, A. Construing scientific evidence: The role of disciplinary knowledge in reasoning with and about evidence in scientific practice. In *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*; Fischer, F., Chinn, C.A., Engelmann, K., Osborne, J., Eds.; Routledge: London, UK, 2018; pp. 66–86.
8. Chinn, C.A.; Brewer, W.F. An empirical test of a taxonomy of responses to anomalous data in science. *J. Res. Sci. Teach.* **1998**, *35*, 623–654. [CrossRef]
9. Ludwig, T.; Priemer, B.; Lewalter, D. Assessing secondary school students' justifications for supporting or rejecting a scientific hypothesis in the physics lab. *Res. Sci. Educ.* **2019**, *51*, 1–26. [CrossRef]
10. Meister, S.; Krell, M.; Göhner, M.; Upmeier zu Belzen, A. Pre-service biology teachers' responses to first-hand anomalous data during modelling processes. *Res. Sci. Educ.* **2020**, *52*, 1–21. [CrossRef]
11. Göhner, M.; Krell, M. Preservice science teachers' strategies in scientific reasoning: The case of modeling. *Res. Sci. Educ.* **2020**, 1–20. [CrossRef]
12. Crujeiras-Pérez, B.; Jiménez-Aleixandre, M.P. Students' progression in monitoring anomalous results obtained in inquiry-based laboratory tasks. *Res. Sci. Educ.* **2019**, *49*, 243–264. [CrossRef]
13. Hug, B.; McNeill, K.L. Use of first-hand and second-hand data in science: Does data type influence classroom conversations? *Int. J. Sci. Educ.* **2008**, *30*, 1725–1751. [CrossRef]
14. Johnstone, A.H.; Sleet, R.J.; Vianna, J.F. An information processing model of learning: Its application to an undergraduate laboratory course in chemistry. *Stud. Higher Educ.* **1994**, *19*, 77–87. [CrossRef]
15. Chinn, C.A.; Brewer, W.F. The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Rev. Educ. Res.* **1993**, *63*, 1–49. [CrossRef]
16. Limón, M. On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learn. Instr.* **2001**, *11*, 357–380. [CrossRef]
17. Mason, L. Responses to anomalous data on controversial topics and theory change. *Learn. Instr.* **2001**, *11*, 453–483. [CrossRef]
18. Lin, J.-Y. Responses to anomalous data obtained from repeatable experiments in the laboratory. *J. Res. Sci. Teach.* **2007**, *44*, 506–528. [CrossRef]
19. Chinn, C.A.; Malhotra, B.A. Children's responses to anomalous scientific data: How is conceptual change impeded? *J. Educ. Psychol.* **2002**, *94*, 327–343. [CrossRef]
20. Hemmerich, J.A.; Van Voorhis, K.; Wiley, J. Anomalous evidence, confidence change, and theory change. *Cognit. Sci.* **2016**, *40*, 1534–1560. [CrossRef]
21. Jeong, H.; Songer, N.B.; Lee, S.-Y. Evidentiary Competence: Sixth Graders' Understanding for Gathering and Interpreting Evidence in Scientific Investigations. *Res. Sci. Educ.* **2007**, *37*, 75–97. [CrossRef]
22. Kang, S.; Scharmann, L.C.; Noh, T. Reexamining the role of cognitive conflict in science concept learning. *Res. Sci. Educ.* **2004**, *34*, 71–96. [CrossRef]
23. Brewer, W.F. Perception is theory laden: The naturalized evidence and philosophical implications. *J. Gen. Philos. Sci.* **2015**, *46*, 121–138. [CrossRef]
24. Chinn, C.A.; Brewer, W.F. Models of data: A theory of how people evaluate data. *Cognit. Instr.* **2001**, *19*, 323–393. [CrossRef]
25. Roberts, R.; Johnson, P. Understanding the quality of data: A concept map for 'the thinking behind the doing' in scientific practice. *Curriculum J.* **2015**, *26*, 345–369. [CrossRef]
26. Duncan, R.G.; Chinn, C.A.; Barzilai, S. Grasp of evidence: Problematizing and expanding the next generation science standards' conceptualization of evidence. *J. Res. Sci. Teach.* **2018**, *55*, 907–937. [CrossRef]
27. Chan, C.; Burtis, J.; Bereiter, C. Knowledge building as a mediator of conflict in conceptual change. *Cognit. Instr.* **1997**, *15*, 1–40. [CrossRef]
28. Osborne, J.F.; Patterson, A. Scientific argument and explanation: A necessary distinction? *Sci. Educ.* **2011**, *95*, 627–638. [CrossRef]
29. Kane, J.E.; Webster, G.D. Heuristics and Biases That Help and Hinder Scientists: Toward a Psychology of Scientific Judgment and Decision Making. In *Handbook of the Psychology of Science*; Feist, G.J., Gorman, M., Gorman, M.E., Eds.; Springer Publishing Company: New York, NY, USA, 2013; pp. 437–459.
30. Sander, E.; Jelemenská, P.; Kattmann, U. Towards a better understanding of ecology. *J. Biol. Educ.* **2004**, *40*, 119–123. [CrossRef]
31. Ladle, R.J.; Gillson, L. The (im)balance of nature: A public perception time-lag? *Publ. Underst. Sci.* **2009**, *18*, 229–242. [CrossRef]
32. Cuddington, K. The "Balance of Nature" Metaphor and Equilibrium in Population Ecology: Biology and Philosophy. *Biol. Philos.* **2001**, *16*, 463–479. [CrossRef]
33. Korfiatis, K.J.; Stamou, A.G.; Paraskevopoulos, S. Images of nature in Greek primary school textbooks. *Sci. Educ.* **2004**, *88*, 72–89. [CrossRef]
34. Zimmerman, C.; Cuddington, K. Ambiguous, circular and polysemous: Students' definitions of the "balance of nature" metaphor. *Publ. Underst. Sci.* **2007**, *16*, 393–406. [CrossRef]
35. Potvin, P. The Coexistence Claim and Its Possible Implications for Success in Teaching for Conceptual "Change". *Eur. J. Sci. Math. Educ.* **2017**, *5*, 55–66. [CrossRef]

36. Smith, R.L. *Ecology and Field Biology*, 5th ed.; HarperCollins: New York, NY, USA, 1996.
37. Sandoval, J. Teaching in subject matter areas: Science. *Annu. Rev. Psychol.* **1995**, *46*, 355–374. [CrossRef]
38. Roth, W.-M.; Bowen, G.M.; McGinn, M.K. Differences in graph-related practices between high school biology textbooks and scientific ecology journals. *J. Res. Sci. Teach.* **1999**, *36*, 977–1019. [CrossRef]
39. Masnick, A.M.; Klahr, D.; Knowles, E.R. Data-driven belief revision in children and adults. *J. Cognit. Dev.* **2017**, *18*, 87–109. [CrossRef]
40. Nitz, S.; Meister, S.; Schwanewedel, J.; Upmeier zu Belzen, A. Kompetenzraster zum Umgang mit Liniendiagrammen: Ein Beispiel für Diagnostik im Lehr-Lern-Labor. *MNU J.* **2018**, *6*, 393–400.
41. Meister, S.; Zimmerman, C.; Upmeier zu Belzen, A. Visualizing pre-service biology teachers' conceptions about population dynamics in ecosystems. *Sci. Educ. Rev. Lett.* **2018**. [CrossRef]
42. Upmeier zu Belzen, A.; Engelschalt, P.; Krüger, D. Modeling as Scientific Reasoning—The Role of Abductive Reasoning for Modeling Competence. *Educ. Sci.* **2021**, *11*, 495. [CrossRef]
43. Sale, J.E.; Lohfeld, L.H.; Brazil, K. Revisiting the quantitative-qualitative debate: Implications for mixed-methods research. *Qual. Quant.* **2002**, *36*, 43–53. [CrossRef] [PubMed]
44. Holmqvist, K.; Nyström, M.; Andersson, R.; Dewhurst, R.; Jarodzka, H.; Van de Weijer, J. *Eye Tracking: A Comprehensive Guide to Methods and Measures*; OUP Oxford: Oxford, UK, 2011.
45. Hartmann, S.; Upmeier zu Belzen, A.; Krüger, D.; Pant, H. Scientific reasoning in higher education. *Z. Psychol.* **2015**, *223*, 47–53. [CrossRef]
46. Mayring, P. *Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution*; Theoretical Foundation, Basic Procedures and Software Solution: Klagenfurt, Austria, 2014.
47. Shaffer, D.W.; Ruis, A.R. Epistemic network analysis: A worked example of theory-based learning analytics. In *Handbook of learning analytics*; Lang, C., Siemens, G., Wise, A.F., Gasevic, D., Eds.; Society for Learning Analytics Research: Edmonton, AL, USA, 2017; pp. 175–187.
48. Roberson, D.B. *Test-Wiseness and Background Knowledge: Their Relative Contributions to High Test Performance*; Mississippi State University: Mississippi State, MS, USA, 2020.
49. Ampatzidis, G.; Ergazaki, M. Challenging Students' Belief in the 'Balance of Nature' Idea. *Sci. Educ.* **2018**, *27*, 895–919. [CrossRef]
50. Shtulman, A. Confidence without Competence in the Evaluation of Scientific Claims. In Proceedings of the Annual Meeting of the Cognitive Science Society, Portland, OR, USA, 11–14 August 2010; Volume 32.
51. Masnick, A.M.; Morris, B.J. Investigating the development of data evaluation: The role of data characteristics. *Child Dev.* **2008**, *79*, 1032–1048. [CrossRef] [PubMed]
52. Berland, L.K.; Lee, V.R. Anomalous graph data and claim revision during argumentation. In Proceedings of the ICLS 2010 Conference Proceedings—9th International Conference of the Learning Sciences, Chicago, IL, USA, 29 June–2 July 2010; Volume 2, pp. 314–315.
53. Fischer, F.; Chinn, C.A.; Engelmann, K.; Osborne, J. (Eds.) *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*; Routledge: New York, NY, USA, 2018.
54. Mahler, D.; Bock, D.; Bruckermann, T. Preservice Biology Teachers' Scientific Reasoning Skills and Beliefs about Nature of Science: How Do They Develop and Is There a Mutual Relationship during the Development? *Educ. Sci.* **2021**, *11*, 558. [CrossRef]
55. Beniermann, A.; Mecklenburg, L.; Upmeier zu Belzen, A. Reasoning on Controversial Science Issues in Science Education and Science Communication. *Educ. Sci.* **2021**, *11*, 522. [CrossRef]

Article

Describing the Development of the Assessment of Biological Reasoning (ABR)

Jennifer Schellinger ^{1,*}, Patrick J. Enderle ² , Kari Roberts ³ , Sam Skrob-Martin ¹, Danielle Rhemer ¹ and Sherry A. Southerland ¹

¹ School of Teacher Education, Florida State University, 1114 W Call St, Tallahassee, FL 32306, USA; sks14b@my.fsu.edu (S.S.-M.); dvandezande@fsu.edu (D.R.); ssoutherland@admin.fsu.edu (S.A.S.)

² Department of Middle and Secondary Education, Georgia State University, Atlanta, GA 30302, USA; penderle@gsu.edu

³ Center for Integrating Research and Learning, National High Magnetic Field Laboratory, Tallahassee, FL 32310, USA; kari.roberts@magnet.fsu.edu

* Correspondence: jls09h@fsu.edu

Abstract: Assessments of scientific reasoning that capture the intertwining aspects of conceptual, procedural and epistemic knowledge are often associated with intensive qualitative analyses of student responses to open-ended questions, work products, interviews, discourse and classroom observations. While such analyses provide evaluations of students' reasoning skills, they are not scalable. The purpose of this study is to develop a three-tiered multiple-choice assessment to measure students' reasoning about biological phenomena and to understand the affordances and limitations of such an assessment. To validate the assessment and to understand what the assessment measures, qualitative and quantitative data were collected and analyzed, including read-aloud, focus group interviews and analysis of large sample data sets. These data served to validate our three-tiered assessment called the Assessment of Biological Reasoning (ABR) consisting of 10 question sets focused on core biological concepts. Further examination of our data suggests that students' reasoning is intertwined in such a way that procedural and epistemic knowledge is reliant on and given meaning by conceptual knowledge, an idea that pushes against the conceptualization that the latter forms of knowledge construction are more broadly applicable across disciplines.

Keywords: scientific reasoning; biological reasoning; assessment; three-tiered assessment; Assessment of Biological Reasoning

Citation: Schellinger, J.; Enderle, P.J.; Roberts, K.; Skrob-Martin, S.; Rhemer, D.; Southerland, S.A. Describing the Development of the Assessment of Biological Reasoning (ABR). *Educ. Sci.* **2021**, *11*, 669. <https://doi.org/10.3390/educsci11110669>

Academic Editors: Moritz Krell, Andreas Vorholzer and Andreas Nehring

Received: 14 July 2021

Accepted: 6 October 2021

Published: 21 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Enhanced learning in science moves beyond memorization and recitation of fundamental concepts to encompass a much larger collection of sense-making activities that resemble the cognitive, procedural, epistemic and social work of scientists [1–3]. Although investigative activities occur in science classrooms in myriad ways, they often limit or even neglect to deeply engage students in the explanatory and evaluative spheres of the scientific enterprise that are essential to the development of scientific understandings [3–6]. Greater emphasis on engaging students in practices reflecting the investigative, explanatory and evaluative spheres of science require supporting students in understanding not only the conceptual elements involved but also the procedural and epistemic function of such practices [3,7]. Such learning not only helps students participate in the development of evidence-based arguments, explanations and models, but also helps them learn to evaluate the quality of different elements of these products and how the processes involved in developing them connect with each other [8,9]. Thus, science learning grounded in these practices also necessitates engaging students in various forms of scientific reasoning where they connect these different activities and products of science in complex yet coherent

ways [10]. We define scientific reasoning as the process that encompasses “the skills involved in inquiry, experimentation, evidence evaluation and inference that are done in the service of conceptual change or scientific understanding” [11] (p. 172), a process that brings together conceptual (i.e., content), procedural and epistemic aspects of knowledge [3,12].

Research on learners’ engagement in scientific reasoning activities demonstrates the complexity of such processes, particularly as they engage in the evaluative and explanatory aspects of science [3,12]. A multitude of factors can shape how students engage in scientific practices that serve as manifestations of scientific reasoning. Scholars have pointed to the need to create time and space in classrooms where students are afforded opportunities to develop the epistemic agency required to engage in reasoning activities to construct knowledge [13]. Research into students’ participation in episodes of critique highlight structural and dialogical elements of argumentation activities reliant on scientific reasoning [14]. Studies of instruction centered around students developing and refining scientific models demonstrate that the concepts that serve as the cognitive objects involved in their reasoning must have a robust quality before students can connect them to broader conceptual elements of models [15]. Enhancing students’ reasoning using scientific models requires engaging their creativity, while also supporting their ability to understand the multiple goals that models can help achieve [16]. However, it is important to note that students’ proficiency with procedural aspects of scientific reasoning, including experimentation and data analysis, are supportive of their learning of conceptual objects and epistemological characteristics [17–19]. There is some evidence to suggest that the cognitive and motivational characteristics of students are also predictive of their ability to reason across broader disciplinary contexts [5].

Much of the research into students’ scientific reasoning when engaged in scientific practices involves intensive qualitative analytical approaches that rely on products resulting from relevant activities [15,20,21]. The composition and quality of students’ arguments [21], models [15,16] and constructed responses to open-ended questions [20,22] can be coded by multiple raters to inductively develop thematic findings or deductively assess the alignment of students’ products to theoretically derived frameworks. Such analyses can be extended by or complemented through separate explorations of students’ reasoning as they are engaged in various types of individual interviews, which are then qualitatively coded [15,23]. Other researchers explore students’ reasoning in action, relying on various analytical approaches employing discourse analysis [14] or observation protocols [6,17] that still necessitate qualitative coding or scoring approaches amongst multiple raters.

Another influential aspect of these analytical approaches concerns the conceptual and disciplinary contexts within which they occur. Many of the studies identified remain tied to particular conceptual areas within specific scientific disciplines, including thermal conductivity in chemistry [21], evolutionary theory or genetics in biology [22,23] and carbon cycling and climate change in Earth science [15]. Limited studies exist where researchers have employed more scalable, quantitative instruments that explore connections between students’ scientific reasoning and broader disciplinary contexts [5], multiple reasoning competencies and skills that can be employed internationally [24,25] and measure competencies among various age ranges [26]. Assessments that do exist have been criticized because they are not psychometrically sound [26]. Additionally, most large-scale measures across various disciplines remain focused on students’ conceptual understanding, limiting the inferential capacity of such work to gain understanding about students’ scientific reasoning [27–30]. Thus, measuring students’ scientific reasoning across a discipline and across dimensions of scientific reasoning through more scalable quantitative approaches remains an ongoing challenge for science education research, something that limits the research that can be conducted.

In light of these challenges, this study focuses on the iterative development and validation of a multiple-choice instrument using qualitative and psychometrically sound quantitative approaches aimed at assessing dimensions of students’ scientific reasoning across ten focal topic areas within biology, entitled Assessment of Biological Reasoning

(ABR). As part of a broader study exploring the influence of teachers sustaining productive classroom talk on student sensemaking [31], the effort described here involved adapting a previously used measure of students' ability to construct scientific explanations through two-tiered, open-ended questioning [17]. Using the instrument and previously analyzed student response data, we developed a three-tier multiple choice assessment exploring each biological topic through a conceptually oriented first tier, a procedural explanatory second tier and a newly developed epistemic third tier exploring students' reasoning supporting their scientific explanations. The study presented below was guided by the following research questions:

What does a three-tier multiple choice assessment measure about students' scientific reasoning across a variety of scenarios relying on fundamental biology concepts?

What are the affordances and limitations of using this approach to measure students' scientific reasoning in biology?

2. Literature Review

2.1. Scientific Reasoning

Scientific reasoning, a central feature of scientific sensemaking, has suffered from the absence of a coherent definition. Early conceptualizations of scientific reasoning present reasoning as a process by which one can develop understandings of science by controlling variables and making causal inferences based on the outcomes of those tests [32,33]. This model, which closely aligns with one methodological approach of science, that of controlled experimentation, represents an overly narrow view of science [34] and does not capture the complex set of reasoning strategies encompassed in the coordination of theories (prior knowledge and beliefs) and evidence needed to generate new knowledge [35,36].

The examination of how these strategies interact and inform one another requires that one engages in the investigative, evaluative and explanatory spheres of science described by Osborne [37] as the spheres that position students to address questions such as "What is nature like?", "Why does it happen?", "How do we know?" and "How can we be certain?" (p. 181). As students engage in exploring these questions, they make observations to understand natural phenomena and to figure out why something happens by constructing and testing models and explanatory hypotheses through empirical investigations and/or data collection that serves as a basis for argumentation and critique, a process by which students consider explanations, the strength of those explanations and how those explanations are supported by evidence [38,39]. When students come to interact in all aspects of these spheres, they are positioned to better engage in a more holistic representation of reasoning which includes conceptual (i.e., content), procedural and epistemic aspects of knowledge [3,12].

Discussions as to whether such reasoning is broadly applicable across domains or is domain-specific exist. Shavelson [40] argued that scientific reasoning can be used when considering everyday decisions. Chinn and Duncan [41] argue that such applicability can be applied to evaluate the trustworthiness of claims about larger scientific issues (e.g., global climate change) presented by the scientific community. These arguments connect with ideas that many of the reasoning aspects, such as a claim that must be supported by evidence, occur across disciplines (e.g., history and literature).

Other argue that scientific reasoning is domain-specific. Samarapungavan [42] suggests that epistemic reasoning is tied to the role of evidence (i.e., what counts as evidence in a knowledge claim, to what extent does it count and why does it count) in bridging conceptual knowledge with practice within a specific disciplinary context. Kind and Osborne [12] describe conceptual (i.e., content), procedural and epistemic aspects of scientific reasoning to require domain-specific concepts or the ontological entities of a discipline to answer questions about "What exists?", the procedures and constructs that help establish knowledge claims and answer causal questions about "Why it happens?" and the epistemic constructs, values and applications that support the justification of these knowledge claims to answer questions about "How do we know?" (p. 11).

Whether domain specific or broadly applicable, scientific reasoning that connects conceptual, procedural and epistemic aspects of knowledge push against the traditional focus in K-12 education of correctly reciting information about content [43]. Instead, by emphasizing these forms of knowledge, students are asked to demonstrate an understanding of content in ways that integrate how they know what they know. For example, when the object of reasoning is to understand whether species are living or nonliving things (i.e., conceptual), students must understand criteria for separating these species (i.e., procedural) and they must understand the role that categorization serves in identifying distinguishing characteristics of living from nonliving things and the particular constructs needed to explain the phenomena (i.e., epistemic [12]).

2.2. Reasoning in Practice

Curriculum and instruction in recent years have focused on positioning students to engage with the forms of knowledge construction involved in reasoning through such activities as model-based and argumentation-driven inquiry. Zagori et al. [15] and others [44–48] suggest that models serve as tools for reasoning because they are developed based on prior knowledge, they are used to make prediction and to generate scientific explanation about how and why a phenomenon works, they are informed based on data collected through investigations and observations and they serve as artifacts of new understandings when the initial model is evaluated and revised. Zagori and her colleagues [15] conducted a quasi-experimental comparative study to understand how modeling-enhanced curricular interventions supported students' model-based explanations (e.g., conceptual understanding and reasoning). They found that students had statistically significant gains in their model-based explanations about water and geosphere interactions as measured through a pre- and post-unit modeling task when supported with a rigorous curricular intervention that provided opportunities for students to engage in scientific modeling practices (intervention 2) compared to an intervention that provided only pre- and post-unit supplementary lessons and tasks involving modeling (intervention 1). These findings were based on a quantitative score for each student across five epistemic features of modeling, including components (i.e., model elements), sequences (i.e., component relationships), mapping (i.e., relationship of model to the physical world), explanatory process (i.e., the connections articulated between cause and effect of system processes) and scientific principle (i.e., connections to underlying scientific theory). When examining these results further, the researchers noted that the features of components and explanatory processes explained the difference in the aggregated feature scores. While the scores of these particular features helped explain students' gains in model-based explanations, they provided less insight into how students themselves conceptualized these and other features in their models. To further understand the results, the researchers examined students' scores on the components and explanatory process features in conjunction with student interview data. One key finding from this examination was that students' models served as reasoning tools to explain how and why water flows underground when students' models included hidden elements under the subsurface of the earth. Swartz and colleagues [47], similarly, found that models can serve as reasoning tools in which students improve their understandings and develop new knowledge that encompasses the explanatory mechanisms and relationships between components of a phenomenon, findings that required the analysis of construct maps and focus group interviews to understand how students construct and use models.

2.3. Assessments of Reasoning

We present these studies not only to acknowledge that efforts are being made in science education to provide opportunities to engage students in scientific reasoning but also to acknowledge the effort and work required to assess students' reasoning capabilities. Such assessments require qualitative examination of student work products (e.g., models, drawings, written work and answers to open-response questions), student interviews, students' discourse and engagement in reasoning tasks and activities [4,14,15,17,23]. Similar

effort and work is required in assessing students' reasoning capabilities in argumentation, the results of which highlight that students often struggle to understand why the construction and generation of claims based on evidence are necessary for science learning [49–51], to analyze and discern quality evidence to substantiate their claims [52,53] and to provide justification for the relationship between claims and evidence to support their argument [50–55].

These findings, while useful in helping us understand students' reasoning capabilities, many of which are tied to specific concepts within a scientific discipline (e.g., groundwater and water systems), are not necessarily sustainable or scalable. In response to issues of scale that go beyond just measuring conceptual understanding, a prominent feature of many large-scale assessments [27–30], instruments to measure students' scientific reasoning have been developed [26]. In a review of 38 test instruments measuring scientific reasoning, Opitz and colleagues [26] found that most tests were related to reasoning skills associated with hypothesis generation, evidence generation, evidence evaluation and drawing conclusions within specific scientific domains, biology being the most common ($N = 13$). They found that newer assessments, those developed from 2002 to 2013 ($N = 27$), measure scientific reasoning competencies as a coordinated set of domain-specific skills as compared to the older assessments ($N = 11$ developed from 1973 to 1989). Additionally, they found that, of the newer assessments, only 17 reported reliability measures and fewer reported validity measures, a finding that led the authors to call the "overall state of psychometric quality checks" unsatisfactory (p. 92). Only 14 of the 38 tests were multiple choice and most were of a closed format following a tiered structure.

Tiered assessments present interconnected questions such as two-tiered assessments that measure content knowledge in tier 1 and related, higher order thinking and explanatory reasoning in tier two [56–59]. For instance, Strimaitis and colleagues [60] developed a two-tiered multiple-choice instrument to measure students' abilities to critically assess scientific claims in the popular media. The 12-item assessment presented students with two modified articles (i.e., dangers of high heels and energy drinks) and asked them to evaluate aspects of the claims presented in each article (tier one) and the logic (tier two) they used to determine their response to tier one. Such tests not only provide opportunities to quantitatively measure students' underlying reasons for their answer choices but they also provide opportunities to assess the alternative conceptions that many students hold related to the particular topic being assessed [61].

While a two-tiered assessment can provide a diagnostic measure of student content knowledge and their explanatory reasoning related to that knowledge, it can suffer from over- or under-estimations of student conceptions [62] or alternative conceptions [63–65], meaning it can fail to differentiate mistakes from such things as lack of knowledge or correct answers due to guessing [66]. To account for these estimation errors, instruments with three and four tiers have been developed. Three-tiered assessments add a third item that provides a measure of the student's confidence in their answer to the first two content and reasoning items [63]. Four tier assessments add additional items to measure the test takers confidence in their prior answers. In a four-tiered assessment, tier one measures content knowledge, tier two measures the student's level of confidence in their answer to tier one, tier three measures reasoning for tier one and tier four measures the student's confidence related to their reasoning in tier three [64]. The inclusion of additional tiers to assess confidence serves as a measure of the student's belief in their own accuracy and provides a level of validity to their answers [67]; however, these tiers do not provide additional measures of a student's higher order reasoning skills nor do they attend to the interrelated conceptual, procedural and epistemic aspects of scientific reasoning that can be difficult to assess quantitatively and are not often assessed in this way.

Informed by the previous work that has been conducted in terms of assessments of students reasoning and motivated by a need for a psychometrically sound measure of students' content knowledge and reasoning skills in biology, this research study focuses

on the development, fine-grained analysis and validation of a multiple-choice instrument aimed at assessing students' scientific reasoning across ten focal topic areas within biology.

3. Methods

This research project is part of a broader professional development study focused on supporting biology teachers' practice to engage students in scientific reasoning through productive scientific discourse [31]. The goal of this assessment is to measure students' explanation of biological phenomena. This assessment was developed based on an existing constructed response assessment used to measure students' conceptual knowledge in biology necessary to evaluate scientific claims [17]. Major concepts in the discipline were selected as foci for the questions, allowing the instrument to serve as an assessment of student learning in both secondary and post-secondary biology courses. The topics that the assessment addresses include cell theory, meiosis, mitosis, photosynthesis and cellular respiration, nutrient cycling, species concepts, evolution and natural selection.

This assessment was designed to understand three dimensions of students' biological reasoning of the 10 focal phenomena listed above operationalized within the four styles of reasoning put forth by Kind and Osborne [12]. These styles include experimental evaluation, hypothetical modeling, categorization and classification, and historical-based evolutionary reasoning, and represent key practices in scientific knowledge generation. Experimental evaluation relates to empirical investigations to establish patterns, differentiate objects and test predictions. Three focal topics fall within this style, including respiration, natural selection and photosynthesis. Hypothetical modeling relates to the construction of models. The focal topics of Mendelian genetics, mitosis and evolution fall within this style. Categorization and classification relate to ordering based on variety and taxonomy. Biological species concept and cell theory align with this style of reasoning. Lastly, historical-based evolutionary reasoning relates to the construction of historical derivations of explanations and development, which include meiosis and nutrient cycling.

The three dimensions of biological reasoning were operationalized within each of these styles of reasoning, including conceptual knowledge (i.e., object of reasoning), procedural knowledge (i.e., use of conceptual knowledge required for reasoning within a specific context) and epistemic knowledge (i.e., ability to justify conclusions based the application of that knowledge). To allow for this structure, each question was framed with an introductory scenario targeting the focal phenomenon with relevant imagery, including graphics, tables, or charts. The first item of the 3-tier question was directed at understanding students' knowledge of specific biological concepts relevant to the focal phenomenon. The second question was aimed at students' use of knowledge, or their application of biological concepts to develop explanations for the focal phenomenon. Finally, the third question asked students to apply reasoning for their explanation by asking them to indicate how relevant biological concepts lead to the explanation of the focal phenomenon. Each tiered question had four answer choices that included a correct choice and distractors, which were developed from expert responses and/or known student responses from previous assessments.

Assessments of this nature should be validated for research purposes with the participant populations that they are intended to be used with. Although multiple views exist on the specific procedures that should be followed for developing educational testing instruments [68–70], a shared consensus suggests that varied pieces of evidence should be collected to demonstrate the properties of an instrument and the validity of the instruments' measurements. Figure 1 provides a graphic identifying the multiple lines of evidence we developed to demonstrate the validity of the ABR. For construct validity, we relied on the input of experts to develop the instrument items and assess how well items measured the targeted, theoretically grounded biological constructs. Experts were comprised of five of the six authors and one high school biology teacher. Of the five authors, three hold two post-secondary degrees in biology and two hold post-secondary degrees in biology and in education. One of the experts holding a post-secondary degree in biology and in education

was also a teacher, represented as teacher #2 in Section 3.1.4. Additionally, the high school biology teacher (Teacher #1), who administered the assessment in her class (discussed in Section 3.1.4), has both a teaching credential and a doctorate in biology. For criterion-related validity, we recruited participants from different populations that theoretically differ in their learning about the focal biological concepts. Finally, we conducted several procedures that improved and demonstrated the reliability of the developed items, including their interpretability by participants, analysis of distractor responses and the internal consistency of the items. We also examined the factor structure of the respondent data to explore how the scores from the instrument should be interpreted.

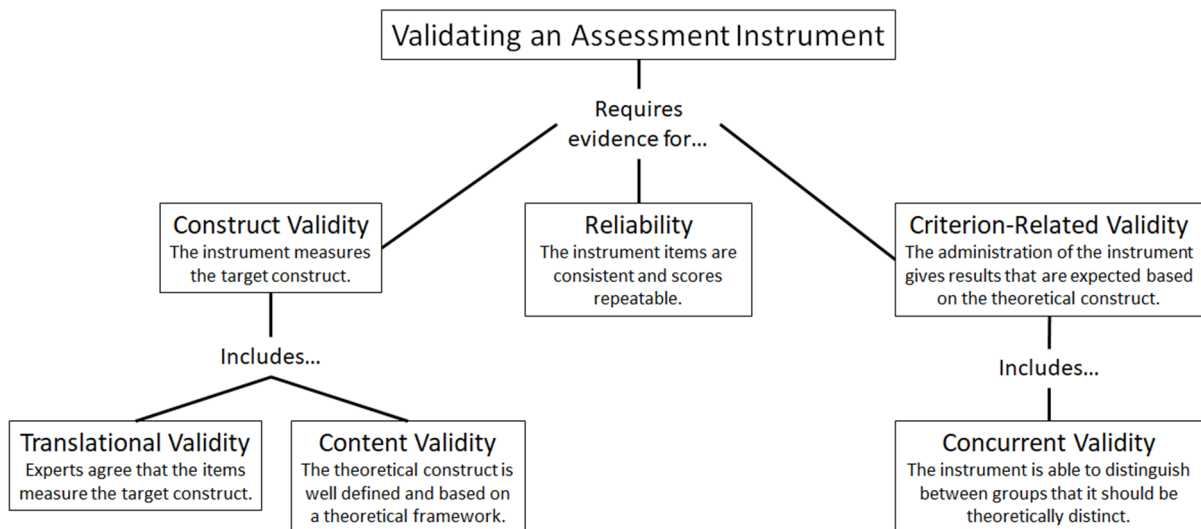


Figure 1. Validation framework used for the Assessment of Biological Reasoning (ABR).

3.1. Data Collection

3.1.1. Exploration of Wording and Coherency Issues

Two rounds of initial testing, including self-recorded read-alouds (see Section 3.1.2) and focus group (see Section 3.1.3) interviews, were conducted to identify possible wording and coherency issues in the 10, three-tiered questions. Each round consisted of a qualitative focus on students' understanding of what each question was asking and the ideas they used to answer the question.

3.1.2. Read-Aloud Interviews

The initial round of testing occurred through self-recorded read-alouds taking approximately from 30 min to 1 h. Seven participants took part in individual read-alouds; one individual had completed high school Biology Honors, three participants were high school biology students and the other three were enrolled in a post-secondary General Biology Laboratory course for non-biology majors. During the read-aloud, each participant read each test item aloud, discussed how they answered the item (e.g., how they arrived at the right choice and why they eliminated certain answer choices), they identified any parts they found difficult or confusing and they made suggestions for item improvement.

3.1.3. Focus Group Interviews

Two virtual focus group interviews were conducted through Zoom. Two participants, one high school Biology and one post-secondary General Biology Laboratory student, took part in the first focus group, which took 2 h and 20 min. During this focus group, the participants answered, annotated and discussed 6 of the 10 questions, including Mendelian genetics, natural selection, nutrient cycling, cell theory, photosynthesis and mitosis. Because of time limitations, the participants in this group answered and made notes on the remaining four questions not addressed in the meeting within one day of the interview. The

second focus group interview had six participants—one high school Biology Honors, two high school Biology and three post-secondary General Biology Laboratory students—and took 2 h and 15 min to conduct. The sequence of the questions was changed for this interview to ensure that feedback for the items that the first group did not have time for in their focus group were examined. In this case, participants answered, annotated and discussed questions related to meiosis, evolution, species concepts and respiration before answering the six questions that the first group started with (i.e., Mendelian genetics, natural selection, nutrient cycling, cell theory, photosynthesis and mitosis).

Both focus group interviews were led by the first author. She followed the same protocol for each interview. In this protocol, participants were introduced to the general structure of the assessment (i.e., three-tiered), they were provided a link to an individual Google document with the assessment questions and then they were asked to work through one three-tiered question individually before coming back together to discuss the question as a group. Students were asked to annotate questions indicating the correct answer, the pieces of the questions that helped them arrive at their answer and to mark any parts that were confusing. Once all students completed the question, the interviewer asked all participants to describe how they solved the problem, the essential pieces of the question that helped them answer it and whether they found any parts of the question or the language of the question difficult, challenging, or confusing. This pattern continued until students cycled through all or most questions. At the end of the interview, participants were asked to discuss if they noticed any changes in how they thought about or read the item for each question and if there were any directions or markers that they wished they had been provided when answering the questions.

3.1.4. Large Sample Data Collection

After completion of the qualitative analysis of the assessment, the instrument was administered to a larger population of students in two rounds to identify if there were any problematic items that potentially needed adjustments. Each round required students to complete the assessment and these data were analyzed for internal consistency.

The first round of analysis focused on examining student assessment data from two teachers (Table 1), one who taught high school Advanced Placement (N = 45 students) and International Baccalaureate Biology classes (N = 15 students) and one who taught a post-secondary General Biology Laboratory course (N = 27 students). The purpose of this analysis was to determine if there were any problematic items and if any items needed to be adjusted. This round also allowed for an examination of item distractors to ensure that they aligned with the internal consistency analyses.

Table 1. Participant information for round 1 data collection.

Teacher	School Type	Course Title	Number of Students
1	high school	Advanced Placement Biology	45
1	high school	International Baccalaureate Biology	15
2	post-secondary	General Biology Laboratory	27

The second round of testing focused on completing final factor analysis, as well as a reexamination of distractor and consistency data. In the distractor analyses, any item with more students selecting an incorrect item choice than the correct item choice was flagged for follow-up review. For internal consistency, Cronbach's alpha was calculated for each scale and alpha values for the scale if each item was removed. We looked for any items where the deletion of the item would increase the scale reliability. More details on the analysis and results can be found in Sections 3.2 and 4. For this purpose, data from three teachers' classrooms collected at the end of the semester were included in the data set. Data were collected from two high school biology teachers (Table 2), one of which participated in the first round of quantitative data collection who taught Advanced Placement (N = 72 students) and International Baccalaureate (N = 20 students) Biology courses and one teacher,

denoted as teacher #3 in Table 2, who taught Advanced Placement Biology (N = 7 students). Additionally, data were collected from teacher #2 who participated in the first round of quantitative data collection. Seven post-secondary students enrolled in her General Biology Laboratory took the assessment in this round.

Table 2. Participant information for round 2 data collection.

Teacher	School Type	Course Title	Number of Students
1	high school	Advanced Placement Biology	72
1	high school	International Baccalaureate Biology	20
2	post-secondary	General Biology Laboratory	7
3	high school	Advanced Placement Biology	7

3.2. Data Analyses

3.2.1. Qualitative Analyses of Individual and Group Interviews

Transcripts were produced for each individual read-aloud and focus group interview for the relevant analyses. For the individual read-alouds, the transcripts were analyzed by the team to identify areas that needed to be clarified, changed, or improved in the assessment. The research team reviewed the participants' responses to the assessment items and the reflective questions concerning clarity of the text and conceptual coherence. As each item and question set was reviewed, the research team identified specific similar challenges mentioned by at least 3–4 students. Similarly, transcripts of the focus group responses for each question set were reviewed. With these transcripts, any issue that maintained the focus of the group's discussion for a significant amount of time was given priority. For the analysis of the individual read-aloud interviews, the research team maintained a stronger emphasis on clarity of the text and how well participants were able to interpret the instrument. For the focus group analysis, greater attention was given to the participants' grasp of the concepts and explanations being provided by the instrument. For both analyses, the researchers collectively identified patterns in the participants' responses and negotiated the manner in which they were addressed as a group. These changes are discussed further in Section 4. Changes occurred after each round of analysis and the revised assessment was used in the next round of data collection.

3.2.2. Quantitative Analyses of Students' Responses to the Instrument

For both rounds of quantitative data analysis presented in this paper, the analyses were conducted using a classical test theory (CTT) approach. The purpose of the first two round of testing for this instrument was to provide preliminary validity evidence before the team conducted large-scale data collection. CTT analyses are more appropriate for smaller sample sizes and provide baseline evidence for the instrument's validity so that the team could begin large-scale data collection for future item response theory (IRT) models with greater confidence in the instrument. The first round of quantitative data was analyzed to assess how well the questions and responses were interpreted by students. For this round of analysis, the research team primarily focused on the distractor analysis and the percentage of students selecting the preferred response. The items that resulted in participants responding with distractor choices for over 50% of the sample were reviewed for clarity. These metrics were determined using SPSS 27. After completion of this analysis, three items were adapted in order to improve performance, where text was altered to clarify distinctions between popular distractor responses and preferred responses. Further, this analysis explored how students in the different courses performed on the assessment to understand the instrument's ability to distinguish between theoretically distinct groups.

The analysis of the second round of quantitative data involved several procedures aimed at assessing the reliability of the instrument as a measure of students' scientific reasoning in biology. The second round of quantitative data analyses focused on the several psychometric properties of the items in the assessment, including item difficulty and discrimination, distractor analysis, internal consistency analysis and exploratory factor

analysis. For distractor analysis, the frequencies of the responses to all four options of each item were calculated using SPSS 27. Any items with distractors which had a higher percentage of students selecting a distractor over the correct answer were flagged for further review. In addition to distractor analysis, we also calculated item difficulty (percentage of students obtaining the item correct or p -value) and item discrimination (a point-biserial correlation between the dichotomous variable for obtaining the item correct and the student's summed score on the rest of the items). The results of the item difficulty and discrimination analyses are presented in Table 3. To evaluate the internal consistency of the instrument, we calculated Cronbach's alpha to measure internal consistency using SPSS 27 for the overall instrument and for each of the tiers in the assessment. Finally, to test for the dimensionality of the instrument, we conducted an exploratory factor analysis (EFA). Dichotomously coded variables were used, with 0 indicating that a student obtained the item incorrect and 1 representing that the student obtained the item correct. The EFA was conducted in Mplus 8.4 [71], using the weighted least squares mean and variance adjusted (WLSMV) estimator.

Table 3. Item Difficulty and Discrimination.

Item Number	Difficulty (p -Value)	Discrimination (Point-Biserial Correlation)
Tier 1		
Q1.1	0.533	0.603
Q2.1	0.598	0.574
Q3.1	0.411	0.599
Q4.1	0.673	0.287
Q5.1	0.411	0.457
Q6.1	0.411	0.428
Q7.1	0.645	0.596
Q8.1	0.626	0.607
Q9.1	0.617	0.515
Q10.1	0.514	0.511
Tier 2		
Q1.2	0.561	0.383
Q2.2	0.579	0.429
Q3.2	0.495	0.292
Q4.2	0.262	0.260
Q5.2	0.514	0.368
Q6.2	0.355	0.233
Q7.2	0.439	0.405
Q8.2	0.673	0.339
Q9.2	0.383	0.282
Q10.2	0.533	0.328
Tier 3		
Q1.3	0.542	0.448
Q2.3	0.607	0.463
Q3.3	0.336	0.406
Q4.3	0.234	0.193
Q5.3	0.477	0.570
Q6.3	0.430	0.432
Q7.3	0.439	0.349
Q8.3	0.589	0.511
Q9.3	0.533	0.469
Q10.3	0.430	0.197

4. Results

4.1. Evidence for Construct Validity—Initial Item Development and Review

As stated previously, the ARB instrument arose from the adaptation of a previously developed and validated instrument aimed at measuring students' ability to construct scientific explanations using core biology ideas [9]. That instrument consisted of two tiers of open-ended, constructed response questions aligned with several theoretical frameworks describing fundamental biological knowledge [10]. This assessment was reviewed by several biologists and biology educators and found to have translational validity, in that all the experts agreed that the instrument measured important concepts and explanations in biology, thus also supporting the construct validity of the ARB. Experts developed ideal answers for the constructed response version that were used to develop the scoring rubrics for the open-ended version of the first- and second-tier questions. For the current ARB instrument, the expert-generated rubrics served as the guide for developing the correct multiple-choice responses for all of the first- and second-tier questions in the ARB. Further, the authentic student responses from data collected in previous studies were reviewed by the research team to develop the distractor responses for the first and second tiers. To establish construct validity for the third-tier questions and responses, a new panel of experts, all who had a minimum of two post-secondary degrees in biology and advanced study in education, reviewed the third-tier questions and agreed they assessed biological reasoning. The third-tier responses also aligned with theoretical descriptions of how core science ideas are used to develop scientific explanations through reasoning [8,72]. Taken together, these efforts support the construct validity for the ARB instrument.

4.2. Evidence for Validity—Outcomes from Qualitative Interview Stages

The analysis of the two rounds of interview data led to several changes in the original iteration of the instrument. One major revision resulting from the initial round of think-aloud individual interviews entailed creating a relatively standardized structure for each tier in the question set for each topic area. The original question stems for the first and second tiers mirrored the question stems from the original constructed response instrument and the third-tier stem followed a general structure of "Which of the following **best describes** your **reasoning** for the choice you made in the previous question (#2)? (2nd tier question)". However, participants experienced difficulty in distinguishing the intent of the third-tier reasoning question from the second-tier question asking them to develop an explanation of the presented scenario using the focal concept from the first tier. Confusion between developing an explanation or the role of evidence in argumentation with the underlying reasoning has been noted in other studies, thus the students' struggle was not surprising [8,73]. To address this issue, all second-tier questions, which originally varied greatly in structure, were aligned more closely to a general form of "Use your knowledge of X (Focal concept in 1st tier) to select the statement that **best explains** Y (Focal scenario for each topic)."

This revised standardized structure was used during the focus group interviews and this set of students described the structure to be clear and logically presented. For instance, they discussed how the first-tier questions required that they pull from their prior knowledge about the concept, the second-tier ones required that they apply that knowledge to a scenario that they considered to have real world applications, which were sometimes novel to them, and the third-tier ones required that they describe their reasoning for that choice. In addition to this group understanding this structure and feeling comfortable in answering the question based on this structure, they also identified that the consistency of this structure helped them understand the nature of the assessment and the connection between the tiers as they progressed through the questions.

Further issues emerging from the analysis of several rounds of interview data broadly related to the semantic structures of the items and potential responses. Several of these issues surfaced as participants considered several of the distractor answer choices. Both individually and in the focus groups, some distractor answer choices seemed too attractive

when compared to desired answer choices. As the research team reviewed these items, the appeal of these distractors followed one of two trends. The first trend involved the distractor response using more generalized language while mainly differentiating through one or two critical terms from the desired response, which typically used slightly more technical wording. The slight variation in ease of comprehension led to the selection of the distractor over the desired response. To address this trend, the responses were edited to limit the level of technicality of each response and to expand the critical elements of the distractor to be more apparent. The second trend in participants' preference for certain distractor responses related to variation in the volume and length of text in the possible responses in the questions. If a particular response was longer and greater in word volume, participants typically deliberated more about their appropriateness and selected those distractors, even if the desired response had less length and volume. To address this trend, the length and volume of all responses for each question set were revised so that they were relatively equal to each other.

One last structural issue that arose for particular question sets involved the nature of the graphics used to accompany the focal scenario for each question set. Specifically, the graphics used in the questions about cell theory, mitosis, photosynthesis and cellular respiration went through several revisions to enhance the clarity of the image and provide a more nuanced representation of the scenario. The photosynthesis and cellular respiration question sets rely on the same experimental scenario using indicators to note the production and use of carbon dioxide in test tubes with plants and animals. The original image used involved black and white graphics only at the beginning of the questions. However, after some revisions, participants engaged in more thoughtful reasoning when color was added to the graphics and the answer choices were aligned to repeated elements from the overarching graphic. As these two questions rely on the evaluation of experimental data, rather than already analyzed forms of data, these revisions appeared to be particularly helpful in supporting participants' engagement with those questions.

4.3. Initial Evidence for Reliability—Outcomes of Quantitative Data Collection and Analyses

For the first round of quantitative data collection, the research team analyzed the results to determine how well the revisions to the textual structure and complexity of the responses supported participants selecting the desired response compared to the distractors. From this analysis, two issues arose that required attention to certain questions and responses. The first issue involved trends in responses to several first-tier questions, which asks respondents to select an answer that best described or defined the focal science concept for the question set. The analysis showed that, for four of these first-tier questions, participants selected one or two distractor responses at levels that were 10–25% greater than the desired response level. Upon review of these first-tier questions, all four followed a similar structure of asking a “negative” question, such as “Select the answer that does NOT represent the products of meiosis.” Based on this pattern in the larger data set, the research team chose to revise those first-tier questions to a more affirmative format, such as “Select the answer that best represents the products of meiosis.” The second issue concerned further challenges involving high similarity between the desired response and a particular distractor for three questions, which were revised further to distinguish between the two selections.

The second round of quantitative data collection provided more participant responses than the first round of data collection, while also allowing all course groups to complete their course of study in biology. The analyses for this data set aimed to explore several psychometric properties of the instrument to provide preliminary evidence for reliability and validity of the instrument. The first analytical step involved further distractor analysis for each item. The results from this analysis demonstrated that only two items had response rates which were significantly higher for a particular distractor (>10%) than the desired response. These particular items included the second- and third-tier questions for the question set involving cellular respiration. For both questions, the more popular distractor

response involved a critical error that misrepresented the role of oxygen in the process of cellular respiration, where O_2 was treated as a reactant rather than a product of the process. Understanding this specific role of oxygen is a key element of a sophisticated understanding of cellular respiration and more advanced reasoning through the experimental scenario presented in the question. Thus, the research team chose to retain these items in their forms as the distractor can help discern learners with more advanced biological reasoning. Only two other distractor responses garnered a slightly higher response rate than their corollary desired response item (<10%), but the review of those items did not demonstrate a compelling need for revision. All other distractors did not reach a response level higher than the desired correct response for the other questions. See Table 3 for a summary of item difficulty and discrimination.

The next psychometric analysis involved assessing the internal consistency of the instrument as a whole and of the three different tiers of question types by calculating a Cronbach's alpha for each subset of the data (see Table 4). For all items together, Cronbach's alpha was 0.905. When looking at the individual tiers within the assessment, the first- and third-tier subsets met the commonly adopted threshold of 0.7 [74]. The second tier had an alpha value slightly below 0.7. Follow-up analyses of item statistics for the second tier showed the deletion of any one item would not have increased the overall internal consistency for this tier, indicating that no item was problematic enough that deleting it from the instrument increased the overall reliability. The reduced internal consistency for the second-tier questions was not unexpected, as these questions are the most unique individually due to the different scenarios presented for each biological topic. Thus, the nature of the appropriate explanations for each scenario involved different reasoning processes, including experimental evaluation, application of analogical models and comparison of classification structures [12].

Table 4. Internal consistency for instrument and question tiers (Cronbach's alpha).

Overall Instrument	1st Tier	2nd Tier	3rd Tier
0.905	0.830	0.672	0.744

To test for initial dimensionality of the instrument, we conducted an EFA in Mplus version 8.4 using the WLSMV estimator. For this, dichotomously coded variables were used, with 0 indicating that a student obtained the item incorrect and 1 representing that the student obtained the item correct. The resulting scree plot is presented in Figure 2. To interpret the scree plot, we first identified the elbow point in the plot, indicating the number of factors at which point factors stop explaining significant portions of the variation and only considered factors to the left of that point significant. Our plot has an elbow point at 2 factors, indicating that only a one-factor model should be considered, based on these data. The plot provides preliminary evidence for a one-factor structure of the item response data. With an elbow point at factor 2, the plot indicates that only the first factor explains a significant amount of variance. For this analysis, two- and three-factor structures were considered. A three-factor structure would be plausible considering the conceptual, procedural and epistemic characters of the different question tiers. A two-factor structure would be plausible in light of the intertwined nature of the procedural and epistemic tiers with respect to the responses. The result of the one-factor structure is intriguing in light of scientific reasoning, as it lends support to the notion that all three elements of reasoning are necessary and possibly inseparable for an instrument in this format. However, we consider these factor analysis results to be preliminary due to the small sample size available. In the future, we plan to distribute the model to a large sample of students and we will conduct a more thorough examination of dimensionality through both exploratory and confirmatory factor analyses as preliminary stages to our planned IRT models.

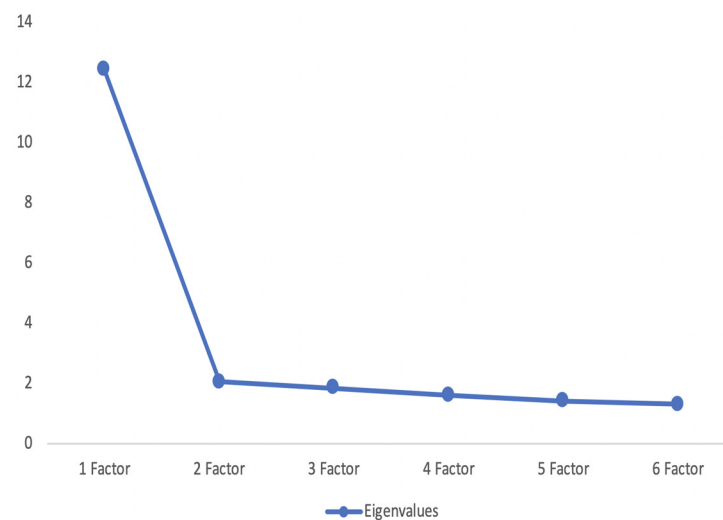


Figure 2. Exploratory factor analysis plot.

4.4. Evidence for Concurrent Validity

Another outcome from the analyses of the larger sets of data is the development of preliminary evidence for the instrument's ability to distinguish between groups of learners who are theoretically distinct. The instrument was administered at the beginning and end of a spring semester course sequence for all three groups. However, the group of participants in the International Baccalaureate Biology course completed a previous semester of biology instruction in the fall. Due to the school schedule structure, a semester-long course in this school equaled what is typically considered a year of typical instruction in most schools. Students in the Advanced Placement Biology and post-secondary General Biology Laboratory course were just beginning their continued study of biology, thus having to rely more on remembered prior knowledge to complete the instrument. That said, the post-secondary students, accepted for study at a research-level university, would be reasonably expected to have at least a slightly more developed conceptual capability in the sciences than Advanced Placement Biology students, who mostly had received introductory-level instruction in life science a few years prior. Thus, it is reasonable to expect the International Baccalaureate Biology students to score better than the other course groups, as they experienced the most recent direct instruction in biology. Further, due to their advanced experience with schooling, it was expected that the post-secondary students would score better than the Advanced Placement Biology students. Disaggregating the data by the different course groups confirmed these expectations, as seen in Table 5, offering evidence to support the concurrent validity of the instrument's ability to distinguish between theoretically different groups.

Table 5. Average percent correct responses across all items by course group, round 1 data.

	Advanced Placement Biology	International Baccalaureate Biology	General Biology Laboratory
Average mean correct	0.35	0.69	0.42
Standard deviation of mean correct	0.10	0.18	0.18

To test for significant differences in the average scores across these groups in the first round of data collection, we ran a Kruskal–Wallis one-way ANOVA in SPSS version 27. The non-parametric Kruskal–Wallis was selected because the data in our sample were not normally distributed, which would have resulted in a violation of assumptions in a traditional one-way ANOVA. The test indicated that, overall, there were significant differences between the groups ($H(2) = 23.130, p < 0.001$). Post hoc tests revealed signifi-

cant differences between Advanced Placement Biology and International Baccalaureate biology ($p < 0.001$) and between International Baccalaureate Biology and General Biology Laboratory ($p = 0.001$), but no significant difference between General Biology Laboratory and Advanced Placement Biology (Table 6).

Table 6. Average percent correct responses across all items by course group, round 2 data.

	Advanced Placement Biology	International Baccalaureate Biology	General Biology Laboratory
Average mean correct	0.43	0.70	0.50
Standard deviation of mean correct	0.10	0.15	0.24

As in round one, we examined the average scores across the different course types to establish concurrent validity for the ARB using the second data set. To test for significant differences in the average scores across these groups, we ran a Mann–Whitney U test in SPSS version 27. The non-parametric Mann–Whitney test was selected for round two data because the sample size for General Biology Laboratory was not large enough to test for statistical significance and the data was not normally distributed, consistently with round one data. The Mann–Whitney test indicated a significant difference between the scores of Advanced Placement and International Baccalaureate Biology ($U = 314.5, p < 0.001$).

5. Discussion

Using the collection of evidence described above, we assert that the preliminary evidence supports the Assessment of Biological Reasoning as a valid assessment instrument for measuring high school students' reasoning capabilities across several major biological topic areas. The resulting ABR assessment consists of 30 questions divided into 10 question sets connected to 10 biological topic areas, with each set including three tiered questions with four answer choices each (see Supplementary Materials for the full instrument). The three-tiered nature of the question sets align with the three recognized dimensions of scientific reasoning [3,12], including a conceptually oriented question comprising the primary object of reasoning, a procedural oriented question that engages the student in developing scientific explanations for the scenarios grounding the question and an epistemically oriented question exploring how a respondent uses the focal science concept to construct their preferred explanatory response. Using a validation framework stemming from the work by Trochim [70] and used in previous validation work by the authors [9], we collected an assemblage of evidence that demonstrates the construct validity, criterion validity and reliability of the ABR instrument.

Through the development of the ABR, the research team gained some insight into the nature of students' reasoning in biology. When developing the instrument, we were not sure if the multiple tiers of questions within a set would be reliant or independent of each other, as each set had a specific focus on a specific ontological/conceptual component but each tier of questions focused on a different component of reasoning. This question regarding the interactive nature of the components stems from descriptions that primarily place domain specificity within the ontological/conceptual component, while the procedural and epistemic components of scientific reasoning have more domain general characteristics [12]. Based on the EFA analysis conducted with the largest sample of responses, the one-factor structure confirmed for the ABR provides preliminary evidence that domain-specific/general distinctions among the three components are not borne out. Rather, although procedural and epistemic dimensions of reasoning may broadly be applied across disciplines, as all science disciplines involve experimental design, modeling and classification, our results suggest that those reasoning components are given meaning by their ontological element. That is, investigating students' ability with certain scientific reasoning activities must pay attention to the ontological/conceptual components of the

activity. This conclusion resonates with other studies that demonstrate that conceptual awareness can improve the overall quality of the verbal argumentation that students engage in, but it is important to indicate that students' epistemic practices can improve separately from conceptual awareness [17]. Additionally, it is important to note that the ABR is mute regarding this point, as the design of the ABR negates this possibly, even if it is sound, given the design of this standardized measure.

The analyses of students' thinking and reasoning during the qualitative data collection also support the intertwined nature of the three tiers of questions within each set. Considering the outcomes described above, an interesting pattern emerged when we examined the questions for which students' expressed difficulties—particularly interpretive difficulties as opposed to simple unfamiliarity with the concept. In the instances, when students encountered interpretive difficulties with a particular question set, we came to understand the students' self-generated descriptions of the focal concepts became a standard by which the students' judged the phrasing of the other response items. It seems that students assessed the language in the responses for the second and third tier through their personal understanding of the focal concepts. This pattern offers an explanation for why the negatively phrased first-tier questions in a previous iteration of the ABR did not produce high correct response rates. This relationship can also help understand how the role of graphics changed and enhanced students' ability to reason through the scenarios, as they provided a conceptual anchor for those questions that could have assisted students in navigating the second- and third-tier questions. The importance of conceptual clarity for respondents' reasoning resonates with findings of earlier studies that speak to the importance of the quality of the cognitive objects involved in students' reasoning [15].

6. Limitations and Implications

The research team recognizes that the ABR instrument and the current validation efforts do have some limitations that should be acknowledged. First, the assessment, while focusing on key biology topics covered in high school and post-secondary education is limited in nature because of this focus. As our results suggest, the ontological/conceptual component of the assessment are interconnected such that the application and reasoning components cannot be disentangled. As such, the ABR instrument is limited in use to biology classes.

Second, the nature of the assessment, while allowing the quantitative assessment of scientific reasoning to be conducted in a controlled format that can be uniformly implemented and easily scored in a short amount of time for a large sample of students, has its limitations [75,76]. One such limitation is that the multiple-choice format is constrained and does not assess reasoning that may occur in what Chinn and Duncan [41] call "the wild". By this they mean that multiple-choice and, even, assessments with open-ended questions do not capture students' reasoning that is observable during performance tasks, inquiry activities, or through direct open-ended, person-centered questioning (questions related to students' ideas) that can be employed by teachers in situ [41,75]. Additionally, while multiple choice tests may have advantages over open-response questions, which often also assess a student's writing ability, they are open to issues of guessing and test taking strategies such as using clues provided by particular words or statements in a question [75].

Third, as a multiple-choice style assessment, there are valid critiques that the wording of response items requires students to comprehend and use language that may not be familiar or representative of their thinking [77]. However, we endeavored to make the language of the response items more accessible by generating many of them from previously recorded student responses and iteratively refining the instrument based on qualitative data from interviews. Similarly, the language used in the question sets is relatively complex and may present challenges for some students. The inclusion of the graphics for each question works to support the interpretability of the questions, but those may not be sufficient and further scaffolding to support students' interpretation of meaning may necessitate further investigation of the ABR. Although the sample sizes for this study were

not overly large, further research being conducted will provide a much larger data set that will help advance the validation of the ABR and the findings related to measuring students' biological reasoning.

Much of the groundbreaking work into students' reasoning in science has been necessarily content-embedded and heavily descriptive, often relying on participant observations and analysis of students' work products and discourse [14,15,20,21]. Given the intensive nature of such investigations, such work is simply not scalable, something that limits the advancement of this line of research. In response to this and to the need for psychometrically sound assessments [26], the ABR represents a contribution to research into secondary students' reasoning in biology, as it is domain- and grade level-specific for measuring students' reasoning in secondary level biology. Although some in-depth assessments of students' reasoning with certain biological topics already exist [22,23], extant assessments across the discipline of biology are primarily limited to measuring conceptual understanding [28]. Thus, the introduction of the ABR represents an advanced tool for the field to use to measure more complex learning and reasoning in secondary biology classrooms, something needed if the field is to move toward larger scale studies involving students' biological reasoning.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/educsci11110669/s1>, Assessment of Biological Reasoning (ABR) Instrument.

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

Funding: This material is based upon work supported by the National Science Foundation under DRL #1720587 and DMR #1644779. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Florida State University (STUDY00001609 approved August 17, 2020).

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.

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

References

1. Crawford, B. From inquiry to scientific practices in the science classroom. In *Handbook of Research on Science Education, Volume II*; Routledge: New York, NY, USA, 2014.
2. Duschl, R. Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Rev. Res. Educ.* **2008**, *32*, 268–291. [CrossRef]
3. Osborne, J. The 21st century challenge for science education: Assessing scientific reasoning. *Think. Ski. Creat.* **2013**, *10*, 265–279. [CrossRef]
4. Walker, J.P.; Sampson, V.; Southerland, S.; Enderle, P.J. Using the laboratory to engage all students in science practices. *Chem. Educ. Res. Pract.* **2016**, *17*, 1098–1113. [CrossRef]
5. Nehring, A.; Nowak, K.H.; Upmeier zu Belzen, A.; Tiemann, R. Predicting students' skills in the context of scientific inquiry with cognitive, motivational, and sociodemographic variables. *Int. J. Sci. Educ.* **2015**, *37*, 1343–1363. [CrossRef]
6. Walker, J.; Sampson, V. Learning to argue and arguing to learn: Argument-Driven Inquiry as a way to help undergraduate chemistry students learn how to construct arguments and engage in argumentation during a laboratory course. *J. Res. Sci. Teach.* **2013**, *50*, 561–596. [CrossRef]
7. Ford, M. Educational implications of choosing “practice” to describe science in the Next Generation Science Standards. *Sci. Educ.* **2015**, *99*, 1041–1048. [CrossRef]

8. Sampson, V.; Enderle, P.; Grooms, J. Argumentation in science education. *Sci. Teach.* **2013**, *80*, 30. [CrossRef]
9. Grooms, J.; Enderle, P.; Sampson, V. Coordinating scientific argumentation and the Next Generation Science Standards through argument driven inquiry. *Sci. Educ.* **2015**, *24*, 45–50.
10. National Research Council. *A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*; National Academies Press: Washington, DC, USA, 2012.
11. Zimmerman, C. The development of scientific thinking skills in elementary and middle school. *Dev. Rev.* **2007**, *27*, 172–223. [CrossRef]
12. Kind, P.; Osborne, J. Styles of scientific reasoning: A cultural rationale for science education? *Sci. Educ.* **2017**, *101*, 8–31. [CrossRef]
13. Stroupe, D.; Moon, J.; Michaels, S. Introduction to special issue: Epistemic tools in Science Education. *Sci. Educ.* **2019**, *103*, 948–951. [CrossRef]
14. Gonzalez-Howards, M.; McNeill, K. Acting with epistemic agency: Characterizing student critique during argumentation discussions. *Sci. Educ.* **2020**, *104*, 953–982. [CrossRef]
15. Zangori, L.; Vo, T.; Forbes, C.; Schwarz, C.V. Supporting 3rd-grad students’ model-based explanations about groundwater: A quasi-experimental study of a curricular intervention. *Int. J. Sci. Educ.* **2017**, *39*, 1421–1442. [CrossRef]
16. Svoboda, J.; Passmore, C. The strategies of modeling in Biology education. *Sci. Educ.* **2013**, *22*, 119–142. [CrossRef]
17. Grooms, J.; Sampson, V.; Enderle, P. How concept familiarity and experience with scientific argumentation are related to the way groups participate in an episode of argumentation. *J. Res. Sci. Teach.* **2018**, *55*, 1264–1286. [CrossRef]
18. Koerber, S.; Osterhaus, C. Individual differences in early scientific thinking: Assessment, cognitive influences, and their relevance for science learning. *J. Cogn. Dev.* **2019**, *20*, 510–533. [CrossRef]
19. Reith, M.; Nehring, A. Scientific reasoning and views on the nature of scientific inquiry: Testing a new framework to understand and model epistemic cognition in science. *Int. J. Sci. Educ.* **2020**, *42*, 2716–2741. [CrossRef]
20. Salmon, S.; Levy, S. Interactions between reasoning about complex systems and conceptual understanding in learning chemistry. *J. Res. Sci. Teach.* **2019**, *57*, 58–86. [CrossRef]
21. Sampson, V.; Clark, D. The impact of collaboration on the outcomes of scientific argumentation. *Sci. Educ.* **2009**, *93*, 448–484. [CrossRef]
22. Haskel-Ittah, M.; Duncan, R.G.; Yarden, A. Students’ understandings of the dynamic nature of genetics: Characterizing undergraduate’ explanations for interactions between genetics and environment. *ICBE Live Sci. Educ.* **2020**, *19*, ar37. [CrossRef]
23. To, C.; Tenenbaum, H.; High, H. Secondary school students’ reasoning about evolution. *J. Res. Sci. Teach.* **2016**, *54*, 247–273. [CrossRef]
24. Krell, M.; Mathesius, S.; van Driel, J.; Vergara, C.; Krüger, D. Assessing scientific reasoning competencies of pre-service science teachers: Translating a German multiple-choice instrument into English and Spanish. *Int. J. Sci. Educ.* **2020**, *42*, 2819–2841. [CrossRef]
25. Krell, M.; Redman, C.; Mathesius, S.; Krüger, D.; van Driel, J. Assessing pre-service science teachers’ scientific reasoning competencies. *Res. Sci. Educ.* **2020**, *50*, 2305–2329. [CrossRef]
26. Opitz, A.; Heene, M.; Fischer, F. Measuring scientific reasoning—A review of test instruments. *Educ. Res. Eval.* **2017**, *23*, 78–101. [CrossRef]
27. Barbera, J. A psychometric analysis of the chemical concepts inventory. *J. Chem. Educ.* **2013**, *90*, 546–553. [CrossRef]
28. Garvin-Doxas, K.; Klymkowsky, M.W. Understanding randomness and its impact on student learning: Lessons learned from building the Biology Concept Inventory (BCI). *CBE Life Sci. Educ.* **2008**, *7*, 227–233. [CrossRef]
29. Hestenes, D.; Wells, M.; Swackhamer, G. Force concept inventory. *Phys. Teach.* **1992**, *30*, 141–158. [CrossRef]
30. Pollock, S.J. Comparing student learning with multiple research-based conceptual surveys: CSEM and BEMA. In *AIP Conference Proceedings*; American Institute of Physics: College Park, MD, USA, 2008; Volume 1064, pp. 171–174.
31. Southerland, S.A.; Granger, E.; Jaber, L.; Tekkumru-Kisa, M.; Kisa, Z. Learning through Collaborative Design (LCD): Professional Development to Foster Productive Epistemic Discourse in Science. National Science Foundation, DRL #1720587. 2017. Available online: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1720587. (accessed on 11 June 2021).
32. Inhelder, B.; Piaget, J. *The Growth of Logical Thinking: From Childhood to Adolescence*; Parsons, A.; Milgram, S., Translators; Basic Books: New York, NY, USA, 1958. [CrossRef]
33. Zimmerman, B.J. Attaining Self-Regulation: A Social Cognitive Perspective. In *Handbook of Self-Regulation*; Boekaerts, M., Pintrich, P.R., Zeidner, M., Eds.; Academic Press: San Diego, CA, USA, 2000; pp. 13–39.
34. Mayr, E. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*; Harvard University Press: Cambridge, MA, USA, 1982.
35. Kuhn, D. *Education for Thinking*; Harvard University Press: Cambridge, MA, USA, 2005.
36. Kuhn, D.; Dean, D. A bridge between cognitive psychology and educational practice. *Theory Pract.* **2004**, *43*, 268–273. [CrossRef]
37. Osborne, J. Teaching scientific practices: Meeting the challenge of change. *J. Sci. Teach. Educ.* **2014**, *25*, 177–196. [CrossRef]
38. Sandoval, W.A. Conceptual and epistemic aspects of students’ scientific explanations. *J. Learn. Sci.* **2003**, *12*, 5–51. [CrossRef]
39. Sandoval, W.; Reiser, B. Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Sci. Educ.* **2004**, *88*, 345–372. [CrossRef]

40. Shavelson, R.J. Discussion of papers and reflections on “exploring the limits of domain-generality”. In *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*; Fischer, F., Chinn, C.A., Engelmann, K., Osborne, J., Eds.; Routledge: New York, NY, USA, 2018; pp. 112–118.
41. Chinn, C.A.; Duncan, R.G. What is the value of general knowledge of scientific reasoning? In *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*; Fischer, F., Chinn, C.A., Engelmann, K., Osborne, J., Eds.; Routledge: New York, NY, USA, 2018; pp. 77–101.
42. Samarapungavan, A. Construing scientific evidence: The role of disciplinary knowledge in reasoning with and about evidence in scientific practice. In *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*; Fischer, F., Chinn, C.A., Engelmann, K., Osborne, J., Eds.; Routledge: New York, NY, USA, 2018; pp. 56–76.
43. Banilower, E.R.; Smith, P.S.; Malzahn, K.A.; Plumley, C.L.; Gordon, E.M.; Hayes, M.L. *Report of the 2018 NSSME+*; Horizon Research, Inc.: Chapel Hill, NC, USA, 2018.
44. Jackson, S.L.; Stratford, S.J.; Krajcik, J.; Soloway, E. Making dynamic modeling accessible to precollege science students. *Interact. Learn. Environ.* **1994**, *4*, 233–257. [CrossRef]
45. Penner, D.E. Complexity, emergence, and synthetic models in science education. In *Designing for Science*; Psychology Press: Mahwah, NJ, USA, 2001; pp. 177–208.
46. Sins, P.H.; Savelsbergh, E.R.; van Joolingen, W.R. The Difficult Process of Scientific Modelling: An analysis of novices’ reasoning during computer-based modelling. *Int. J. Sci. Educ.* **2005**, *27*, 1695–1721. [CrossRef]
47. Schwarz, C.V.; Reiser, B.J.; Davis, E.A.; Kenyon, L.; Achér, A.; Fortus, D.; Shwartz, Y.; Hug, J.; Krajcik, J. Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *J. Res. Sci. Teach. Off. J. Natl. Assoc. Res. Sci. Teach.* **2009**, *46*, 632–654. [CrossRef]
48. Berland, L.; Reiser, B. Making sense of argumentation and explanation? *Sci. Educ.* **2009**, *93*, 26–55. [CrossRef]
49. Erduran, S.; Simon, S.; Osborne, J. TAPping into argumentation: Developments in the application of Toulmin’s argument pattern for studying science discourse. *Sci. Educ.* **2004**, *88*, 915–933. [CrossRef]
50. Clark, D.B.; Sampson, V. Personally-seeded discussions to scaffold online argumentation. *Int. J. Sci. Educ.* **2007**, *29*, 253–277. [CrossRef]
51. Jimenez-Aleixandre, M.P.; Bugallo Rodriguez, A.; Duschl, R.A. “Doing the lesson” or “doing science”: Argument in high school genetics. *Sci. Educ.* **2000**, *84*, 287–312. [CrossRef]
52. Sandoval, W.A.; Millwood, K.A. The quality of students’ use of evidence in written scientific explanations. *Cogn. Instr.* **2005**, *23*, 23–55. [CrossRef]
53. Osborne, J.; Erduran, S.; Simon, S. Enhancing the quality of argumentation in school science. *J. Res. Sci. Teach.* **2004**, *41*, 994–1020. [CrossRef]
54. Ryu, S.; Sandoval, W. Improvements to elementary children’s epistemic understanding from ssutatin argumentation. *Sci. Educ.* **2012**, *96*, 488–526. [CrossRef]
55. Zohar, A.; Nemet, F. Fostering students’ knowledge and argumentation skills through dilemmas in human genetics. *J. Res. Sci. Teach.* **2002**, *39*, 35–62. [CrossRef]
56. Adadan, E.; Savasci, F. An analysis of 16–17-year-old students’ understanding of solution chemistry concepts using a two-tier diagnostic instrument. *Int. J. Sci. Educ.* **2012**, *34*, 513–544. [CrossRef]
57. Chen, C.C.; Lin, H.S.; Lin, M.L. Developing a two-tier diagnostic instrument to assess high school students’ understanding-the formation of images by a plane mirror. *Proc. Natl. Sci. Counc. Repub. China Part D Math. Sci. Technol. Educ.* **2002**, *12*, 106–121.
58. Griffard, P.B.; Wandersee, J.H. The two-tier instrument on photosynthesis: What does it diagnose? *Int. J. Sci. Educ.* **2001**, *23*, 1039–1052. [CrossRef]
59. Treagust, D.F. Development and use of diagnostic tests to evaluate students’ misconceptions in science. *Int. J. Sci. Educ.* **1988**, *10*, 159–169. [CrossRef]
60. Strimaitis, A.M.; Schellinger, J.; Jones, A.; Grooms, J.; Sampson, V. Development of an instrument to assess student knowledge necessary to critically evaluate scientific claims in the popular media. *J. Coll. Sci. Teach.* **2014**, *43*, 55–68. [CrossRef]
61. Tan KC, D.; Taber, K.S.; Goh, N.K.; Chia, L.S. The ionisation energy diagnostic instrument: A two-tier multiple-choice instrument to determine high school students’ understanding of ionisation energy. *Chem. Educ. Res. Pract.* **2005**, *6*, 180–197. [CrossRef]
62. Chang, H.P.; Chen, J.Y.; Guo, C.J.; Chen, C.C.; Chang, C.Y.; Lin, S.H.; Su, W.J.; Lain, K.D.; Hsu, S.Y.; Lin, J.L.; et al. Investigating primary and secondary students’ learning of physics concepts in Taiwan. *Int. J. Sci. Educ.* **2007**, *29*, 465–482. [CrossRef]
63. Caleon, I.; Subramaniam, R. Development and application of a three-tier diagnostic test to assess secondary students’ understanding of waves. *Int. J. Sci. Educ.* **2010**, *32*, 939–961. [CrossRef]
64. Caleon, I.S.; Subramaniam, R. Do students know what they know and what they don’t know? Using a four-tier diagnostic test to assess the nature of students’ alternative conceptions. *Res. Sci. Educ.* **2010**, *40*, 313–337. [CrossRef]
65. Peşman, H.; Eryılmaz, A. Development of a three-tier test to assess misconceptions about simple electric circuits. *J. Educ. Res.* **2010**, *103*, 208–222. [CrossRef]
66. Hasan, S.; Bagayoko, D.; Kelley, E.L. Misconceptions and the certainty of response index (CRI). *Phys. Educ.* **1999**, *34*, 294–299. [CrossRef]
67. Renner, C.H.; Renner, M.J. But I thought I knew that: Using confidence estimation as a debiasing technique to improve classroom performance. *Appl. Cogn. Psychol. Off. J. Soc. Appl. Res. Mem. Cogn.* **2001**, *15*, 23–32. [CrossRef]

68. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. *Standards for Educational and Psychological Testing*; American Educational Research Association: Washington, DC, USA, 2014.
69. Kline, P. *The Handbook of Psychological Testing*, 2nd ed.; Routledge: New York, NY, USA, 2000.
70. Trochim, W.M. *The Research Methods Knowledge Base*, 2nd ed.; Atomic Dog: Cincinnati, OH, USA, 1999.
71. Muthén, L.K.; Muthén, B.O. *Mplus: Statistical Analysis with Latent Variables: User's Guide*; Version 8; Muthén & Muthén: Los Angeles, CA, USA, 2017.
72. McNeill, K.L.; Krajcik, J. Inquiry and scientific explanations: Helping students use evidence and reasoning. *Sci. Inq. Second. Setting* **2008**, *121*, 34.
73. McNeill, K.; Knight, A. Teachers' pedagogical content knowledge of scientific argumentation: The impact of professional development on K-12 Teachers. *Sci. Educ.* **2013**, *97*, 936–972. [CrossRef]
74. Taber, K.S. The use of Cronbach's alpha when developing and reporting research instruments in science education. *Res. Sci. Educ.* **2018**, *48*, 1273–1296. [CrossRef]
75. Harlen, W. *Assessment & Inquiry-Based Science Education: Issues in Policy and Practice*; Global Network of Science Academies: Trieste, Italy, 2013.
76. Simkin, M.G.; Kuechler, W.L. Multiple-choice tests and student understanding: What is the connection? *Decis. Sci. J. Innov. Educ.* **2005**, *3*, 73–98. [CrossRef]
77. Lee, O.; Quinn, H.; Valdes, G. Science and language for English language learners in relation to Next Generation Science Standards and with Implications for Common Core State Standards for English Language Arts and Mathematics. *Educ. Res.* **2013**, *42*, 223–233. [CrossRef]

Article

Models as Epistemic Artifacts for Scientific Reasoning in Science Education Research

Marvin Rost ^{1,*}  and Tarja Knuuttila ² ¹ Austrian Educational Competence Centre Chemistry, University of Vienna, 1090 Vienna, Austria² Department of Philosophy, University of Vienna, 1010 Vienna, Austria; tarja.knuuttila@univie.ac.at

* Correspondence: marvin.rost@univie.ac.at; Tel.: +43-1-4277-60353

Abstract: Models are at the core of scientific reasoning and science education. They are especially crucial in scientific and educational contexts where the primary objects of study are unobservables. While empirical science education researchers apply philosophical arguments in their discussions of models and modeling, we in turn look at exemplary empirical studies through the lense of philosophy of science. The studied cases tend to identify modeling with representation, while simultaneously approaching models as tools. We argue that such a dual approach is inconsistent, and suggest considering models as epistemic artifacts instead. The artifactual approach offers many epistemic benefits. The access to unobservable target systems becomes less mysterious when models are not approached as more or less accurate representations, but rather as tools constructed to answer theoretical and empirical questions. Such a question-oriented approach contributes to a more consistent theoretical understanding of modeling and interpretation of the results of empirical research.

Keywords: science education; scientific reasoning; models and modeling; philosophy of science

Citation: Rost, M.; Knuuttila, T. Models as Epistemic Artifacts for Scientific Reasoning in Science Education Research. *Educ. Sci.* **2022**, *12*, 276. <https://doi.org/10.3390/educsci12040276>

Academic Editors: Moritz Krell, Andreas Vorholzer and Andreas Nehring

Received: 14 December 2021

Accepted: 8 April 2022

Published: 13 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Imagine a chemistry teacher trying to explain the volume contraction that occurs when water and ethanol are mixed using the famous demonstration of mixing the corresponding volumes of lentils and beans. Since the contraction of the liquid mixture is a non-trivial consequence of a change in hydrogen bonding length and is not mechanistically explainable by smaller molecules that fill the gaps between larger molecules, the lentil-bean demonstration is clearly misleading. Moreover, another major source of confusion is also simultaneously introduced: molecules are identified with solid spheres while imposing the same identification on single atoms. How is a learner supposed to know when it is appropriate to apply such a structural simplification of volume contraction?

One would expect scientists to be prepared to point out the analogies and simplifications used in the bean-lentil model by stating the assumptions involved. Yet, presenting such assumptions is not a trivial task. Not only should the empirical researchers be able to articulate their own theoretical framework, e.g., psychological constructs or observational premises. Moreover, they would simultaneously need to refer to the specific subject on which, e.g., learning groups acquire knowledge or skills. Instead of such explication work, a representational perspective is often adopted, where the notion of representation is, implicitly or explicitly, understood as a structural or other kind of similarity relation between a model and its supposed target system. But how is one able to understand the lentil-bean example according to such a representational notion of modeling? Indeed, the representational approach cannot easily make room for the fact that models are intrinsically tied to human made inferences, actions, and interpretations, and not just to the natural objects, processes, and systems they study.

Apart from chemistry education e.g., [1–4], the representational approach to modeling is also widely present in other fields of science education research e.g., [5–7]. In its reliance on the representational approach to model-based reasoning, science education research

does not differ too much from the mainstream philosophical discussion of modeling (e.g., Weisberg [8]). However, one peculiarity of the science education research literature studied in this article is that the representational conception of modeling features in them side-by-side with the notion of models as tools. By contrast, in philosophical literature, approaching knowledge and human action from the perspective of tool use has traditionally been used to criticize the representational conception of knowledge [9,10]. Another idiosyncrasy of science education research is its tendency to move in between scientific models and students' supposed mental models, as if they were comparable entities. Such an understanding of model-based reasoning has its advantages, for example, in zooming in on the subject matter in question and the students' understanding of it, yet it turns out to be highly problematic in practice.

An unreflective use of the notions of a model and representation, causes problems both in empirical research and in classrooms. If empirical researchers try to discuss their studies within their research communities without properly laying out the assumptions underlying their respective understanding of models, especially when studying scientific reasoning processes, they run the risk of losing common ground, on something that has empirically been observed to be the case [11,12]. Confusions ensue, not because the researchers would have conducted erroneous experiments or miscalculated their statistics, but rather because the results arrived at are not on par with the underlying theoretical assumptions concerning modeling. Likewise, if science teachers are using models merely as representational depictions free from ontological and other assumptions—and not as tools for addressing, e.g., a particular scientific question—it may cause confusion in their learning groups. Such confusions may arise even if every part of the lesson was correct in view of the content to be taught, as well as regarding the level of knowledge of the students.

Given the centrality of the notion of a model in both research and teaching [13–15], we call for a more coherent and explicit treatment of it. With such a theoretical goal in mind, we will argue for the artifactual approach to models [16,17], through presenting and analyzing exemplary empirical and theoretical studies from the field of science education research. The artifactual account approaches models as concretely built artifacts that are constructed by employing various kinds of representational tools. Central for the epistemic functioning of models, according to the artifactual account, is their constrained design that facilitates the study of particular theoretical and empirical questions, and learning from models through their construction and manipulation [18].

In what follows, we study some exemplary studies on modeling within the field of science education research, discussing their degree of internal consistency regarding their respective theoretical frameworks and empirical findings. We then present the artifactual notion of models, and conclude our paper with suggestions on how to think about modeling as a question-oriented activity that employs concrete artifacts for scientific reasoning. Such an artifactual perspective, we claim, can lead to better practice, and stronger mutual understanding within the field.

2. Model-Based Reasoning in Empirical Science Education Research

2.1. *Scientific Reasoning in General*

Empirical studies in science education research discuss scientific reasoning in various ways. Scientific reasoning is often typically linked to (formal) argumentation and delineated between the theoretical extremes of domain-generality and domain-specificity [19]. Such a middle-ground between the domain-generality and domain-specificity appears well-justified. If, on one hand, scientific reasoning were necessarily tied to specific domains, a general path of doing science would be blocked. On the other hand, too exclusive an attention to domain-generality could lead to general theorizing with (nearly) no contact to domain-specific knowledge. This is often the case with many modeling endeavors that apply cross-disciplinary model templates, such as various network models, to different, often distant domains [20,21].

In addition to the domain-specific dimension of reasoning, the empirical literature has identified general patterns within reasoning processes on the basis of interviewing

researchers about their work, or empirically testing learning environments [22–24]. Such patterns of reasoning are not bound to specific subjects [25–27], but are instead hypothesis-driven, and supposed to work iteratively. They are usually implemented as follows: first, a question is elicited in a research-oriented learning environment and a preliminary hypothesis is formed; second, a suitable scientific investigation is planned and conducted; third, the collected observational or experimental data is processed and referred back to the prior hypothesis; followed finally by the assessment of the hypothesis with respect to the original question, generating new questions and hypotheses, and leading to an iterative process of inquiry.

The aforementioned patterns emerge from different, subject-oriented studies in science education cf. [28–30]. They range from kindergarten [31] and preschools [32,33] to higher education [34]. Given the vast diversity of these implementations, one may ask whether there is a generalizable perspective from which scientific reasoning skills, e.g., formulating adequate questions for respective investigations, could be approached in learning and teaching sciences. One such perspective is provided by model-based reasoning.

2.2. Model-Based Scientific Reasoning

Models are an active area of research within science education research. A host of different perspectives on models and modeling have been introduced and further developed, starting from a focus on visualization [35], to presenting a broad, comprehensive overview of different perspectives on modeling [14].

A substantial part of the discussion of models and modeling in the literature on scientific reasoning aims at straddling the divide between general modeling methods and subject-specific applications. In this regard, models have often been considered as mental or abstract entities, that express formal relations between propositions [36,37], as heuristic devices serving to generate concrete analogies [38], or connecting disciplinary knowledge to data, thus generating explanations [39]. When turning to the generalization-oriented end of the field, assessments of competencies with regard to model-based reasoning [12,40,41] focus on the reasoning processes of learners. As such, the role of models as tools for reasoning within research processes is understood as competency-based cf. [42], and it is presently under vast empirical investigation, since it relates closely to international educational standards, thus shaping the teaching and learning of science.

Within science education research, the notion of “model-being” has offered a prominent approach to the ontology of models [43]. This approach draws together a collection of different perspectives, incorporating also considerations from the philosophy of science, and providing the foundation for the competence model of model competence [44,45]. The related epistemological notion of models is agent-based [46,47]. The agent-based perspective addresses the circumstances in which a model is referred to as such: who, where, when, and to what end does a human judge an object as being a model [48,49]? Despite several empirical educational studies e.g., [50–52], the understanding of models in science [53] and science education [11] remains diverse. Such diversity in understanding has led to an astonishing [44] as well as surprising [43] diversity in model classification schemes. It is, therefore, crucial to further examine the concepts, terminologies, and differentiations native to science education in order to pave the way for a more unified analysis of models and model-based reasoning in science education research [54].

2.3. Examples from Science Education Research

In this section, we will discuss the incoherent treatment of models in science education research, using empirical examples. We begin by presenting two detailed cases, followed by shorter analyses as well as a discussion of a well-received theoretical approach to models. On the basis of our observations on these studies, we call for a more consistent use of the artifactual notion of models. The studies chosen are exemplary in that they are careful in articulating how they understand the notion of a model, and modeling as a particular kind of theoretical reasoning. However, their conclusions seem partially inconsistent in that their

theoretical starting points do not necessarily align with their empirical findings, a problem that we trace back to the authors' representational stance towards models and modeling.

2.3.1. Models as Generative Tools

Schwarz et al. [5] provide an interesting case of a partially inconsistent treatment of models in that they argue for understanding models as generative tools at the level of their empirical analysis, yet defining models in a more traditional representational and abstract way. The authors report a learning progression among primary and middle school students where the more sophisticated way of using and understanding models is to view them as tools that "[...] can support [the students'] thinking about existing and new phenomena." (p. 640), instead of understanding models as literal illustrations of what a single phenomenon is like. At the higher end of this progression, students are able to construct multiple models of related phenomena and appreciate their respective advantages and weaknesses.

Similarly, Schwarz et al. elaborate on students' metamodeling [27,55,56] knowledge: the ability of the learner to elucidate inconsistencies which, in turn, can help her and her teacher productively intervene in learning processes, e.g., by turning the inconsistencies into starting points for conceptual change [57]. Such metamodeling knowledge concerns the learner's understanding of models and modeling in science, and progresses from considering models as "[...] good or bad replicas of the phenomenon [...]" (p. 647) to that of viewing them as explanatory and changeable tools, whose changes are crucial for developing new questions. The same goes for the elements of scientific practice, i.e., what learners actually do within the boundaries of their tasks [41,58,59].

In spite of their practice-oriented approach to models as tools, Schwarz et al. define a model as "[...] an abstract, simplified, representation of a system of phenomena that makes its central feature explicit and visible and can be used to generate explanations and predictions." (ibid. p. 633). Moreover, the authors distinguish models from other representations:

"It is important to clarify that not all representations are models. Models are specialized representations that embody aspects of mechanism, causality, or function to illustrate, explain, and predict phenomena." (ibid. p. 634).

In referring to the function of models, the authors ascribe to the agent-based account of models (to be discussed more in detail in the sections below). Consequently, it is the users' judgment about the proper means to serve a particular purpose that is crucial for something to function as a model. Yet, at the same time, the authors still hold on to the realist [60] understanding of models as objective representations of systems/phenomena. Moreover, Schwarz et al. assess the students' success in terms of what they think about the respective phenomena, leading to the question of whether the modeling activity would not be considered successful if a phenomenon were not recovered correctly. But the correctness of the students' supposed mental content would be hard to assess if, say, the targeted system in question were on a submicroscopic level. Or, alternatively, would the modeling activity be successful if a learner "[...] consider[ed] how the world could behave according to various models" (ibid. p. 640)?

The definition of models proposed by Schwarz et al. tries to bridge the gap between models as representations and models as tools, while in their empirical study the students' progression clearly proceeds from naive realist correspondence between a model and a phenomenon towards more reflective uses of models as tools for scientific reasoning. Moreover, their notion of models as abstract representations of phenomena does not seem to suit the concrete examples of the models produced by the students in the empirical study. It is these concrete models rendered by different representational means – pictures, symbols, and language – that researchers focus on (in addition to students' commentary) and not any abstract mental models within students' heads.

On the one hand, Schwarz et al. consider representing, or rather depicting, phenomena, and iteratively revising for better or alternative explanations and predictions as a central

defining aspect of a successful modeling cycle (“elements of practice”). On the other hand, the authors also refer to models as means of eliciting What If? questions (“metaknowledge”). These aspects are not mutually exclusive. However, without explicating the connection between realistically conceived representational aspects of models, and the progression towards a more instrumentalist understanding of them, the epistemological stance of the authors remains unclear. Finally, the authors treat both visible (e.g., a shadow emerges), as well as non-perceivable (e.g., particle movement) target systems, as representable on the same scale. It appears to us that these problems concerning the interpretation of their empirical study are due, at least in part, to the unexplained, and to some extent inconsistent, notion of models with which the authors operate. While we have thus detected inconsistencies between the different parts of the study of Schwarz et al., we wish to emphasize that we do not contest their empirical study or the learning activity reported, but rather the concessions that their instrumental view on modeling nevertheless makes to representational realism.

2.3.2. Model-Based Reasoning and NOSI Views

As a second example, we analyze a study from chemistry education research [61] that attempts to link a three-dimensional framework of scientific reasoning competencies (i.e., observing as theory-driven activity, experimenting as manipulation of variables, and using models as tools for inquiry) with views on the nature of scientific inquiry (the so-called NOSI views). Models are important for testing “[...] hypotheses about an original object [...]” (ibid. p. 2720). The reference to original objects is crucial for the authors’ definition of models:

“The model serves as [a] substitute object [...] for an original object when these objects are not available—due to ethical or practical reasons, for example. Students use the model not only to derive a hypothesis or to explain a phenomenon but also to derive data about the original object with regard to their research questions. They test models against data on the underlying original object and reflect the validity of their assumptions.” (ibid. p. 2719)

We would like to highlight that Reith and Nehring simultaneously present models both as tools, i.e., human-shaped constructs, and as surrogates for non-perceivables, i.e., structural representations. Similarly to Schwarz et al., this conflation results in an inconsistent view on models. Reith and Nehring claim that a “naive view” on models considers a model “as an exact copy of reality” (ibid. 2720). Such a view supposes that a surrogate could directly represent atomic features, e.g., by using lentils and beans. An “informed view”, in contrast, “[...] [carries] out investigations on models. [Scientists] test hypotheses about an original object using models” (ibid. p. 2720). However, the authors do not explicitly delineate the circumstances under which a model object is a mere copy of reality (i.e., a direct representation), or when to refer to it as an appropriate tool to represent assumptions about a target system. Moreover, we wish to point out that introducing models as surrogates for original objects, such as assumed submicroscopic entities, runs the risk of reifying these entities in principle, thus falling back on a naive view time and again. Such a view would make the example of mixing legumes as a representation of the respective submicroscopic system to learn something about volume contraction irrelevant at best. The vegetables can hardly represent smaller/larger molecules with regard to canonical mechanistic explanations, i.e., changes in hydrogen bond length. The artifactual notion of models does not start from assuming such a possibility of direct representation, thus lifting the argumentative burden when it comes to the supposed structure of non-perceivables. However, if a teacher would like to introduce how scientific modeling works, surrogate reasoning on the basis of the simplified legumes-molecules correspondence does not add value to the learning environment unless this correspondence is further elaborated. In such a case, understanding the hypothetical nature of the model would be the very point of the exercise. If the same teacher would like to convey canonical knowledge about how molecules are supposedly structured, then the lentil-bean demonstration is inappropriate,

given the numerous and partly contradictory portrayals of submicroscopic entities in, e.g., chemistry textbooks. With this in mind, it would be helpful if science education researchers, exemplified by Reith and Nehring as well as our other cases, refrained from constituting their understanding of modeling via a dyadic relation between models and target systems. We will elaborate the artifactual alternative in the respective sections.

2.3.3. Further Studies on Modeling

The works of Schwarz et al. and Reith and Nehring provide examples of the many cases within the field of science education research where, in our view, more consistency in how models are approached and defined would have strengthened their educational implications. In this section, we give a brief overview of some other studies, representative of the current state-of-the-art in the field of science education. What they have in common is that they tend to take a largely unarticulated representational stance towards models, while simultaneously treating models as tools. A more reflective and differentiated approach that pays heed to different kinds of representational tools and their epistemic affordances would have been more appropriate. Such an approach would help addressing, e.g., the difficulties science learners face in acquiring generalizable knowledge when they are confronted with symbolic abstract representations that are presented as mere surrogates for unobservables (e.g., particles, forces or pedigrees) [62,63].

Cheng et al. [6] present models as epistemic tools “[...] to represent [students’ and teachers’] ideas, or to coherently explain the mechanisms underlying target events.” (2019, p. 5). The notion of a model as an abstract representation seems to provide purchase both to students’ and teachers’ ideas and to the real-world target systems. Abstraction plays a crucial role in both cases, as it allows treating the subjects’ ideas as mental models, as well as scientific models as abstract theoretical representations of mechanisms underlying the phenomena. However, a mental model of a theoretical idea and the allegedly correct representation of a submicroscopic target event are two different things. Additionally, if models are considered as abstract representations, why would a student be assessed as a more advanced modeler if she were able to visualize submicroscopic mechanisms, i.e., sketching what is considered a structurally correct depiction of magnetic field lines? In our view, this would testify to the students’ ability to employ cultural representational tools correctly, which is not accounted for when models are conceived of as abstractions.

Luca and Zacharia [64] neither clearly distinguish the students’ supposed mental models from models of external real-world target systems, nor pay due attention to the importance of the external representational tools with which models are constructed. They point out that “[...] models can be both concrete and conceptual (i.e., models we create in our mind) in nature, in our case we refer to external/physical models.” (p. 195). Yet in their discussion of model construction, students are supposed to “[...] mentally bring the model’s content/elements together in order for the model to take shape (have a structure). This cognitive process takes place immediately before learners start constructing their concrete artifacts/models.” (ibid.). Consequently, models reduce to the “[...] externalization of the components and underlying mechanism of a phenomenon/system”. How did the students have access to the underlying mechanics of a phenomenon/system in the first place? Only by collecting observations and experiences, as Louca and Zacharia seem to suggest? This question becomes all the more puzzling as the authors judge the accuracy of a model in terms of how well it represents the features of a respective phenomenon. The study focuses on phenomena at the macroscopic level, yet purports to apply to the representation of the underlying (unobservable) mechanisms as well.

Likewise, when turning to chemistry-focused studies, the question of how a learner could gain competency in handling the problem of unobservable structures remains challenging. Stieff et al. [7] work on what they label as concrete molecular models, i.e., three-dimensional ball-and-stick objects for grasping spatial structures of submicroscopic targets. The authors stress the importance of the empirical investigation of representational competence, which they measure by a test of translating between different chemical depictions of molecules, e.g., translating from the Newman projection to the Fischer projection.

However, such an approach already presupposes a structurally adequate relation between the projections and their respective target systems and thus, elucidates how the participants are able to express and communicate certified knowledge about the atomic scale (ibid., p. 345).

Oliva et al. [65] studied the competence of modeling among secondary students learning about chemical change. Various kinds of representational tools were used (fruits and bowls, Lego pieces, balls of plasticine, discs of colored cards, etc.) “[...] as mediators between the students’ intuitive understanding and school science models.” (p. 751). The authors used several different qualitative and quantitative methods of data analysis. They delineated modeling as an activity employing a range of inferential and reasoning processes that require the students to be able to “[...] interpret, handle, and express phenomena and situations using as certain variety of signs, whether propositional or iconic in format [...]” (p. 753). In their analysis, Oliva et al. tend to conflate mental models and scientific models, in that they relate the students supposed “intuitive models” to “school science models” implying that the application of the same notion of a model to both enables their comparison. Moreover, despite their attention to actual representational tools, they invoke a meta-representational perspective to draw together and evaluate multiple representations. Yet, they do not explicitly attempt to state the conditions under which such an evaluation would be judged to be adequate or successful. Provided that Oliva et al. also subscribe to the models-as-tools approach, it would have been advantageous to address the contributions of different kinds of representational tools in producing scientific understanding as well, rather than focusing only on their supposed unification at the meta-representational level.

2.3.4. *Models of* and *Models for*

The theoretical discussion of models within science education research attempts to navigate between models as tools and models as representations, but not always entirely consistently. Gouvea and Passmore [47] make a distinction between *models of* and *models for*, following Fox-Keller [66], who views models in molecular biology as tools for both theoretical reflection and instruments for material intervention. Gouvea and Passmore argue that “[...] the models of account [of models] often comes alongside models for, which makes it seem like an alternative on equal footing” (ibid. p. 57). They are critical of such attempts, advocating for approaching scientific models as tools for understanding, explanation, and prediction, especially in classroom settings. In their view, the *models of* accounts “[...] are less able to support students’ epistemic agency in doing science because they tend to treat models as representations of what is known rather than as tools to be used in generating new knowledge.” (ibid, p. 50).

Although Gouvea and Passmore are focusing on science classrooms, they also put forth a more general agent-based conception, inspired by the pragmatic accounts of scientific representation within philosophy of science. While we find their agent-based conception of modeling interesting, and also deserving of philosophical attention, some clarification of what they mean by representation would be needed. However, despite their stated intention of approaching models primarily as tools, “i.e., models for a purpose”, their model appears to take the “representational axis” of *models of* on par with the “epistemic axis” of *models for* cf. [46]. As a consequence, the authors distinguish the representational relationship between a model and “a phenomenon”, from the understanding of seeking questions and other epistemic aims of the model. To be sure, Gouvea and Passmore underline that the “[...] two axes are interdependent and inform and constrain each other.” (ibid. p. 53). The epistemic agents, in their view, “[...] specify how models will represent phenomena [...]” (ibid.), while the representational axis concerns the “[...] respects and degrees the model represents the features of some phenomenon.” (ibid.). Yet, given that they do not explicate the notion of representation, it is difficult to tell what they in fact are committed to concerning the representational axis of their account. Gouvea and Passmore claim that their agent-based conception of a model is based on the work of Suárez [67,68] and Giere [46], but these pragmatic accounts would not separate the representational axis from the epistemic axis. Instead, the epistemic aims of the model users are an integral

part of Suárez's and Giere's analyses of representation (i.e., the "representational axis" of Gouvea and Passmore).

In order to see what is at stake more clearly, in the next sections we will provide a brief overview of the philosophical discussion of models and representation. This overview is followed by our suggestion as to how the artifactual account of models as tools should be framed, such that it does not get subsumed by the representational account. Two things are especially important in this regard. First, although models are constructed by using representational tools, the systems specified by these tools do not need to accurately represent any real-world target system. They can also compose fictional, or merely hypothetical systems, addressing various possibilities and impossibilities [69]. Second, the crucial challenge for any account that seeks to approach models as tools is to explain how they could provide scientific understanding without falling back on the representationalist assumption that they do so in virtue of representing some real-world target system more or less accurately.

The artifactual account seeks to account for these challenges by focusing on the scientific and empirical questions models are constructed to answer, instead of supposing that models would need to have any determinable and fixed relationship to some real-world target system. From this perspective, *models of* are *models for*.

3. Contemporary Philosophical Perspectives on Models

As we have discussed above, there appears to be a tension in the science education literature about whether to consider models as tools or representations. The studies discussed above treat models as tools while simultaneously adhering to an unexplained notion of representation. This bifold strategy tends to lead to incompatibilities at both the theoretical and empirical levels. That is precisely what the artifactual account of modeling aims to avoid.

We have found that while the notion of models as epistemic tools has gained traction in science education research [55,70], the notion has also been used inconsistently. However, the problems involved do not certainly concern just science education researchers. They are present also in those contemporary philosophical accounts of models and representation that approach the epistemic value of modeling in terms of representation, yet also invoke pragmatic aspects, i.e., factors relating to the use of models (e.g., [71,72]).

For example, Chakravartty [72] distinguishes between the informational and functional dimensions of modeling. The functional dimension of models refers to their capacities to support scientific reasoning, while the informational dimension relies on representation, conceived loosely as some kind of similarity between a model and its target system. Accordingly, the functional dimension presumes the informational dimension. Chakravartty asks: "how [...] could such [inferential and reasoning] practices be facilitated successfully, were it not for some sort of similarity between the representation and the thing it represents—is it a miracle?" (ibid. 201). We suspect that the same kind of reasoning motivates the attempt of science education researchers to merge the notion of models as tools with the idea of representation: if the world behaves as if it were made of invisible particles, why not accept the inference to the best explanation (and the world it depicts)?

The question posed by Chakravartty is thorny indeed as we will discuss in the next sections, and yet, it quite obviously tends to put the cart before the horse. At least when it comes to scientific practice, models are frequently tools for probing what kinds of systems and causal processes might bring about particular kinds of phenomena. Consequently, they are tools for finding out what might be the case instead of representing what is known to be the case (though successful models may gain the status of certified knowledge over time).

3.1. Perspectives on Representation

The idea that modeling has something to do with representation has a long history within philosophy of science, yet Suárez [73] finds out that "the modeling attitude" of both the British (e.g., Thomson and Maxwell) and German scientists and philosophers (e.g., Helmholtz, Hertz, and Boltzmann) of the 19th century, were in fact nuanced. Apart from relying on similarity, resemblance, and analogy, the scientists in question were acutely

aware, according to Suárez, about the relativity of knowledge. Boltzmann's Encyclopaedia Britannica entry, "Models", is especially interesting in this regard [74]. On one hand, he writes about models as "representations in thought" and on the other, he invokes the material and tangible objects that scientists have created for assisting their thoughts.

This practice-oriented tradition of considering models as concrete things or their mental images later on became entangled with the semantic and syntactic conceptions of theories with their notion of a model derived from mathematical logic. The resulting "model muddle" [75], does not, however, mean that the notion of a model itself would be vague cf. [47]. Rather, the word model is used in various ways, in different contexts. As our focus is on science education, we limit ourselves to those philosophical discussions that explicitly concern models in scientific practice. Two contemporary discussions are of special interest in this regard: the pragmatic accounts of representation, and the accounts of modeling that instead of concentrating on the representational relationship, address model construction. The latter accounts study how scientists learn from building and manipulating hypothetical systems, frequently called models, without supposing that such model systems would accurately reproduce some features of some target systems of interest.

3.1.1. The Pragmatic Account of Representation

The pragmatic accounts of representation aim to provide an alternative to the so-called substantive accounts of representation. Such substantive accounts—i.e., structural or other less formal similarity accounts—seek to explain how models give us knowledge by asking how a model represents its target system. The answer is provided by the relationship between the constituent parts and relations of the model and those of its supposed target system. In other words, such accounts analyze representation in terms of a structural, or some other kind of similarity relation, between the model and its target. Yet, the structuralist and similarity accounts of representation have been rather conclusively criticized within the recent philosophy of science discussion: they have been found lacking when it comes to both their logical and practical dimensions [76,77]. As a result, several structuralist philosophers have attempted to amend their accounts of representation by either accommodating some specific criticisms concerning e.g., the direction of representation [78], or by extending their account of representation by including pragmatic elements with it [79]. On the other hand, many philosophers have increasingly embraced a pragmatic approach to models and representation.

To put it bluntly, the basic issue is this: the pragmatists of scientific representation claim that it is not possible to analyze the representational relationship without making the users and their aims an integral part of it. In terms of Gouveau and Passmore's agent-based conception of models, this would mean that the representational and epistemic axes would coalesce instead of the remaining separate dimensions of modeling. For example, Giere [80] analyzes scientific representation as a four-place relationship: "S uses M to represent W for purposes P", where S is an individual scientist, group of them or a scientific community, M is a model, and W stands for an "aspect of the real world, a (kind of) thing or event." This form can be translated into the following, more informal statement: "Scientists use models to represent aspects of the world for various purposes" (ibid. p. 747). In other words, the users' goals become a part of the definition of representation and as a result, one cannot analyze representation without taking them into account. Suárez [67,68] also grounds his account of representation in the representing activity of modelers. His inferential account of representation has two parts: the *representational force* and the *inferential capacities*.

The representational force of a model is due to the practice of scientists using it as a representation of an intended target. Yet, representational force alone is not enough to make any model a scientific representation. Consequently, in order to function as a scientific representation, the model must possess inferential capacities enabling a competent user to draw valid inferences regarding the target.

What is important, then, to note about the aforementioned pragmatic accounts of representation is their minimal nature: a model represents a target system if it is used to

represent. That in turn, according to Suárez, is based on the inferential capacities of the model, and some norms concerning valid inferences. What those inferential capacities and norms consist of, Suárez does not say. As a result, pragmatists do not say anything substantive about representation, as they do not invoke any deeper constituent relation, such as similarity or structural mapping, between the parts of the model and the parts of the target. What pragmatists are in fact saying is that a model is a representation, if it is used as such. And such a notion of representation does not, by design, explain why models give us knowledge, something that the substantive accounts attempted to do.

The question then becomes: How can one understand how models give us knowledge if representation is trimmed down into such a thin notion that it cannot explain the epistemic productivity of modeling? The answer would need to be sought for somewhere other than from the notion of representation.

3.1.2. Model Construction

Morrison and Morgan [81] focus on learning from constructing models instead of using them as representations. They approach models as investigative instruments, whose construction and manipulation enable scientists to learn from them. They view models, rather than as representations, as mediators between theory and data. Likewise, Weisberg [8] considers models as independent from any uniquely determinable relationships to the worldly target systems (ibid. p. 218). Modeling is for Weisberg an art of *indirect representation*, one of building and studying hypothetical systems that will only be related to some particular real-world systems at a later stage of the modeling cycle, if at all.

Many areas of contemporary modeling testify to such an indirect approach with only a few manifest ties to some clearly identifiable target systems. For example, economics has often been accused of modeling without an attempt to relate the highly abstract models to economic realities [82]. The same kinds of concerns have also been raised in biology [83].

Despite paving a way for understanding models as tools, both Morrison and Morgan, as well as Weisberg, eventually invoke the notion of representation as well. Morrison and Morgan are careful to note, however, that they do not consider representation to be “mirroring” or “correspondence”, yet they do not develop their notion of representation any further. They mainly note that it should be thought of as “[...] a kind of rendering—a partial representation that either abstracts from, or translates into another form, the real nature of the system or a theory, or one that is capable of embodying only a portion of a system. [81], p. 27.” Weisberg [84] formulates a formal account of similarity on the basis of Tversky’s set-theoretic account [85] that has not succeeded to create any noticeable interest in the philosophy of science community.

To sum up, the lively philosophical discussion of modeling and representation has not settled on any one notion of representation. The structuralist and similarity accounts of representation have proven difficult to flesh out in any satisfactory fashion, while the pragmatist accounts have remained overly deflationary. Given these difficulties concerning the notion of representation, the artifactual approach to models builds directly on the idea that models are human-made objects, whose construction and use in scientific practices is the key to their epistemic value.

4. Models as Epistemic Artifacts

Instead of assuming that models more or less faithfully represent real-world target systems, the artifactual account focuses on how models as purposefully designed artifacts provide access to the empirical and theoretical questions scientists are interested in. According to a standard philosophical definition, artifacts are intentionally made or altered objects, whose aim is to accomplish some purpose [86]. Such definition pays heed to (i) the *aim* that an object has in some human practice and (ii) its *intentional production or alteration* that involves the use and modification of various kinds of materials. Consequently, from the artifactual perspective scientific models are human-made objects that are typically designed for answering some pending scientific problems and built by making use of a variety of representational tools (i.e., various symbolic, semiotic, and material resources).

Both of these aspects of model construction—purposeful design and the representational tools employed—are important for how a model can provide access to a problem scientists are dealing with.

4.1. Purposeful Design

The artifactual account envisages models as human-made objects that can have multiple epistemic uses. In science and science education, they can be used for explanatory, predictive, and assessment purposes, for example. Traditionally, especially the explanatory and understanding bearing dimensions of modeling have been accounted for by appealing to representation. Instead of approaching models as representations of real-world target systems, the artifactual account seeks to analyze the epistemic dimension of models through their interrogative function: addressing the scientific questions models are designed to answer. The constrained construction of a model is the key to its interrogative functioning. Models typically consist of a system of dependencies, designed to answer a pending scientific question, motivated by theoretical and/or empirical considerations [18,47,87]. In other words, relevant theoretical and empirical knowledge needs to be built into it, both through its specific construction and the question(s) it addresses.

For example, in constructing his version of the Lotka-Volterra model, Volterra set out to answer the question of whether the variations in the populations of predators and prey could be produced solely by “[...] the purely internal phenomenon, due only to the reproductive power and to the voracity of the species as if they were alone. [88], p. 5.” To study this question, Volterra wrote a pair of nonlinear differential equations concentrating only on the dynamics between two species, one of which preys on the other, while also acknowledging the importance of external causes for the actual fluctuations in populations. Indeed, at the time when he published his results, the fluctuations in predator and prey populations were usually attributed to some external causes [89]. Akerlof’s celebrated model of the “market for lemons” that earned him a Nobel prize provides an example from economics. It studies through a simplified model of used cars the question of how quality uncertainty can lead the bad quality cars to drive out the better quality cars, leading even to market collapse.

What is important to note about both Volterra’s and Akerlof’s models is that they are not inherently tied to any specific target system, but are rather hypothetical systems constructed to study general theoretical questions. The general character of the dynamics they study have allowed for their application to sundry other problems.

Alfred Lotka used the Lotka-Volterra model to study, apart from biological systems, also chemical systems. Later on, the Lotka-Volterra equations were applied across different disciplines to study various kinds of target-systems, ranging from class struggle to models of technology diffusion [20]. Moreover, the Lotka-Volterra equations have been used as a basic simple model to study the complex behavior of nonlinear systems [90]. Akerlof, in turn, did not intend in his classic article to only study markets for used cars. In fact, the market for used cars was for him simply a “finger exercise” chosen for its “[...] concreteness and ease in understanding rather than for its importance or realism” (p. 489). Akerlof’s focus was on the effects of asymmetric information more generally, and in his famous article he proceeds from presenting the model of used cars to study its implications for various, more important topics, such as the health insurance market, the employment of minorities, and credit markets in underdeveloped countries.

The artifactual perspective can better capture the *initial motivation underlying the construction* of such exemplary models as the Lotka-Volterra model and Akerlof model. From the perspective of scientific practice, to which science education naturally relates on a large scale, one of the main problems of the representational approach is due to its basic unit of analysis: the model-target pair.

Viewing models as inherently targeting a particular real-world system leads to problems concerning their accuracy and misrepresentation, but more importantly, misses their most important scientific contributions. Consequently, the artifactual approach focuses on the questions models are designed to address. Due to their interrogative function, models

are already embedded in existing theoretical and empirical knowledge, e.g., knowledge concerning fluctuations in populations, or market failures due to degrading quality of goods offered. Instead of gesturing at (an unexplained notion of) representation, the artifactual account zooms in on model construction and the access it bestows for further scientific theorizing and exploration, including the application of the model to other domains [91].

4.2. Representational Tools

As we have argued above, the way a model is constrained is crucial for its epistemic functioning; striving for accurate representation of some particular target system is frequently less helpful if the goal is to tackle some more general question, as is often the case with modeling. In such tasks, minimal and unrealistic models may be explanatorily useful: such models may isolate some hypothetically relevant, or difference-making features for particular patterns of interest [92–94]. Moreover, the use of mathematical and statistical methods entails simplification and unification as well [95].

In contrast to the representational approach that focuses on the general and abstract features of the relationship of representation, the artifactual approach emphasizes the concrete, workable dimension of models rendered by various representational tools, such as differential equations in the case of the Lotka-Volterra model. The concrete workability of models explains how scientists can learn by building and manipulating them [81]. For instance, Volterra's ability to draw important results from a highly idealized hypothetical system shows that in order for models to be epistemically useful, they do not need to correspond more or less accurately to real-world systems and processes.

This learning process is facilitated through articulating different kinds of relationships within a model with some particular representational tools, concretely manipulating them, and reconfiguring the model in view of further questions. Such work can lead to various kinds of explanations, predictions, and theoretical results, and may contribute to novel experimental designs and the construction of artificial and synthetic systems [96].

The fact that the epistemic importance of the concrete workable dimension of models has not received due recognition can partially be traced back to the tendency of treating models as abstractions. Such a tendency is understandable given the importance of mathematical and computational modeling in contemporary science. Once models are considered as abstract entities, likening them, or at least comparing them, to mental models seems an easy step to take, as we have seen above. Such a step should be resisted, however. The concrete workable dimension of models does not boil down to their material aspects only, it also applies to mathematical modeling as the case of the Lotka-Volterra model shows.

Most of Volterra's papers on biological associations are highly technical mathematics, consisting of the study of the mathematical properties of the Lotka-Volterra model and its variations. In other words, the differential equations provided Volterra the workable dimension of the Lotka-Volterra model, and the study of these equations gave him several results that could be given a biological interpretation. He would not have come up with these results had he simply mentally conceived the predator-prey system: the differential equations provided him a representational tool to access the dynamics between the two populations.

The representational tools employed in modeling typically consist of various symbolic or semiotic devices (mathematical, iconic, diagrammatic etc.) that serve as vehicles for conveying different kinds of content. However, these vehicles need to be embedded in representational media that furnish the material means with which representations are produced and manipulated (such as ink on paper or digital computer in which simulations are run) [18,87,96,97].

The representational media and their materiality play different epistemic roles depending on the type of model in question that has led to the perception that some models, such as mathematical models, are inconcrete, whereas other models, such as scale models, are concrete. But on closer inspection, such a distinction between concrete and inconcrete models tends to lead astray. For instance, there is accumulating evidence that the perceptual and sensorimotor engagement with external mathematical representations is crucial for

mathematical reasoning over and above them functioning as mere scaffolds for mnemonic and communicative tasks [98,99]. On the other hand, the Phillips-Newlyn model, a hydraulic model of a macroeconomy in which colored water flows and accumulates in a system of tanks and channels, does not reduce it to its material embodiment. As such, it would hardly be interpretable as a model, let alone an economic model. Instead, it gives a concrete form to the conceptualization of the economy in terms of stocks and flows that has a long history in economic theorizing [100].

4.3. Representing and Justifying

It may seem puzzling that the artifactual approach seeks to explain the epistemic value of modeling without invoking representation, yet emphasizes the importance of representational tools. No contradiction is involved as representation in the sense of establishing a relationship between a model and a real-world target system should be distinguished from representing something *within* the model.

Representation in this latter sense refers to the use of representational tools to convey some content that is a precondition for claiming any representational relationship between a model and some external target system. Such distinction between these two notions of representation is embedded in the recent philosophical literature, where modeling as an activity of building and studying models is distinguished from establishing a representational relationship between a model and a target system. For instance, Weisberg [8] argues that the practice of indirect representation distinguishes modeling from those theoretical strategies that rely on abstract direct representation. Modeling, Weisberg claims, is engaged in indirect representation as modelers are primarily interested in studying their models, before trying to relate them to some real-world, or merely possible targets. Indeed, apart from providing possible explanations of the actual states of affairs, models also enable inferences concerning unactualized possibilities [87,96,101]. Such modal reasoning constitutes one of the main ways in which models are used in scientific practice [102].

Regarding the modal dimension of modeling, the artifactual account approaches the question of justification through model construction: a model is constructed for the purpose of probing theoretical and empirical consequences. Thereby, it becomes necessary to independently justify any kind of representational relationship (if only because of underdetermination). The fact that some models are used as representations does not provide justification for model-based results in and of itself. Although, part of the justification is already built-in due to the previously established theoretical, empirical and representational resources used in model construction [103]. The already established use of differential equations, and the mechanistic approach of isolating the components and their interactions, in addition to the observations on fluctuations in fish populations were resources already built-into Volterra's model. Due to these pre-established resources and knowledge, the relationship of representation is not pivotal for explaining how models are able to generate knowledge: it is not needed to connect a model to the empirical world as the connection is already partially built-in.

Finally, it goes without saying that in order to establish the external validity of a model, more is needed than consistently analyzing the built-in connection from successfully certified models. Such external validation in work-in-progress models typically proceeds by triangulating different epistemic means: other models, experiments, observations, and background theories. These processes of triangulation are often not easily recognizable due to their complex and indirect nature. Justifying models does, therefore, not happen through individual model-target comparisons as, e.g., the representational approach would have it, but rather by rigorously questioning models, even at the level of research programs, being distributed in terms of time, place, and epistemic labor.

5. Future Challenges and Implications

Equipped with the notion of models as epistemic artifacts, we turn in this section to two concrete examples, where a more theoretically consistent approach to modeling would have strengthened the already valuable educational implications. In our examples,

science education researchers implement straightforward empirical strategies according to the notion of models as artifacts, while such an approach has been less prevalent in philosophy of science. Nevertheless, these science education studies tend to set the concrete representational tools (e.g., sketches) aside, turning to discuss the mental models of students, as if a direct connection between them and the students' concrete modeling products could be established by the researcher in some unproblematic manner. In contrast, and in line with the artifactual approach, we emphasize that science education should focus on the epistemic value of concrete products in investigating the system of interest.

5.1. Model-Based Learning and Reasoning

As an important next step towards a better mutual understanding of model-based learning and reasoning, we propose below how to clarify the connection between the empirical studies' results, and their respective theoretical frameworks by drawing on the insights of this paper.

First, it would be helpful to focus on whether or not a learner was able to refine iteratively, and in a justified manner, concrete model objects (by sketching, modeling clay, etc.). In this regard, the question about the adequate rendering of canonical scientific knowledge appears to be of secondary importance. Yet, such an approach may appear unsatisfying at first: what scientist would give credit to a learner who gives justified, yet evidently false explanations about the behavior of a target system? It might seem that useful representations should not include disproven assumptions, at least within learning environments. Nevertheless, a learner may eventually be able to confront the experienced scientist/teacher with cases where hypothetical speculation is an intrinsic part of daily scientific business. Moreover, the learner may wonder why atoms are described as identical to tiny solid spheres in every introductory chemistry lecture, when the scientific community *knows* that this is not the case. When viewed from the artifactual perspective such assumptions do not appear so baffling, as they highlight the question-oriented character of modeling, providing thus a reasonable, though underappreciated, starting point for science classes [104].

Second, carefully choosing an appropriate target system presents challenges of its own. It does matter whether a learner either works on how introducing a species into a biotope affects the population of another species and comes up with a numerical association by counting and extrapolating, or tries to find a mechanistic explanation of ice maintaining its temperature while melting during heat supply. Both tasks can be approached through modeling, yet they are fundamentally different in terms of their underlying goals, i.e., numerically predicting or mechanistically explaining the target system. The situation gets even more complex if, contrary to the purely predictive goal, one inquires about the mechanisms that lead to the influence of one species on another, e.g., a predator-prey relationship or a displacement of another population due to an advantage in reproduction. Likewise, associating heat supply to state transitions and making predictions without asking for submicroscopic mechanisms is in itself valuable [105], highlighting the paramount importance of the question to be asked for any modeling activity. Thus, it is crucial to explicitly distinguish whether the aim of modeling is to present what is currently accepted as *being the case* in the field [106], pp. 141, or whether the focus is on practicing to think about and test the consequences of *what if something were the case?* [41]

5.2. Models and Subject-Specific Content

Inconsistencies of subject-specific models are rarely explicitly addressed in science education research [107], and, if discussed at all, they are approached within the context of multiple modeling [108–110]. However, presenting to a learner multiple models of a certain target system (e.g., Bohr's model vs. Lewis' structures) does not inform the learner when it is appropriate, e.g., to refer to electrons as particles circling around an atomic core, in contrast to electrons as fixed bonding pairs. The models do not reveal, in and of themselves, to what end and under which circumstances they were constructed, and what seems even worse from the learner's perspective is that they seem not to be true at the same time. In

this regard, learners and teachers alike should be encouraged not to suppose that they could perfectly state how unobservables, or other lesser known phenomena, are structured: multiple models of the same target systems should be regarded as a normal phenomenon in scientific research. Consequently, teachers, learners, and researchers should focus on the modal dimension of modeling, seeking plausible estimations, fruitful depictions, and how-possibly explanations. We were not able to identify such a consistent modal focus within the investigated studies.

We hereby turn to vindicating the lentils and beans model to a certain degree: if students work with this representational vehicle in response to a relevant research question, they can learn about chemistry as a matter of course. For example, if the bean-lentil demonstration was used to explain volume contraction, how could the structural relationship between the demonstration and the target system be justified in the first place? If we did not have any other evidence for such a relationship, we could adopt a question-oriented approach: *what if* the lentils and beans were structurally equal to water and ethanol particles? Subsequently, experiments would come into play and different liquids could be mixed and their behavior documented. Fortunately, in the sense of fostering model-based reasoning, mixtures exist that show a volume expansion, which falsifies the assumption of smaller particles fitting into the gaps between the larger particles as a general principle. That falsification could potentially lead to a more sophisticated reasoning activity that makes use of students' artifacts. These artifacts, in turn, could be integrated into standardizable frameworks under current development, e.g., stepwise procedures for the modeling of target systems in chemistry classrooms [15]. However, teachers and researchers should be careful about their presuppositions of unobservables; which of them appear to be resolved, and which of them side-stepped, via an over-simplified representation. While models as epistemic artifacts are constructed by representing what could plausibly, or possibly, be the case, and are thus able to convey scientific content [111,112] that does not yet justify supposing that they would accurately depict the structure of their target systems—as a representationalist would have it. A little sphere is not structurally equal to a molecule.

6. Conclusions

We have claimed that scientific reasoning can usefully be viewed as a question-oriented investigation. Modeling provides a prime example of such an activity. We have suggested that an explicit and reflective discussion of models as artifacts serves to prevent a relapse into viewing models as straightforward, uniquely determinable representations of target systems. We have observed in science education research a conflation of mutually exclusive epistemological accounts of models and representation, i.e., adhering to both pragmatist and structuralist perspectives. If a researcher refers to models as constructed tools, it is difficult to maintain a representational dyadic model-target relationship as a unit of analysis. Modeling submicroscopic mechanisms for explaining or predicting the behavior of, e.g., chemical target systems is a case in point. As we have shown, straddling between the pragmatist agent-based and the representational similarity-based and structuralist approaches to modeling breeds inconsistencies both on the theoretical level and between the theoretical definitions of models and the interpretation of empirical results.

Consistently understanding and explicating models as artifacts is helpful since it fosters an understanding of science as being revisable by keeping the focus on the interrogative, uncertain, and fallible nature of scientific reasoning. Thus, the studied target systems can be worked on with models as metaphorical magnifying glasses, hammers, or screwdrivers. Consequently, the artifactual approach shifts the focus of the discussion of scientific modeling within science education research from accurate representation into the learning of how to do science. Moreover, since the artifactual approach views any representational relationships between models and some real-world targets as contingent scientific achievements, it prompts researchers and teachers to reflect on the assumptions they make about target systems.

Finally, we find plenty of room for a dialogue between philosophy of science and science education research, a dialogue that is already happening. The link to teaching

makes science education research a worthwhile area of study for philosophers of science: philosophy cannot be considered just a source for trickling down theoretical ideas to empirical sciences. Especially practice-oriented philosophers of science are interested in what scientists think and do to gain knowledge about the world, and for this task, they need case studies and empirical research. Science education researchers are uniquely positioned to do just that: studying and conveying scientific reasoning at different levels of teaching, learning, and researching. Therefore, we advocate a fruitful and critical discussion between philosophers of science and science education researchers concerning their theoretical presuppositions and definitions, addressing also the question of how to plan and/or revise empirical studies on the basis of such reinvigorated mutual understanding.

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

Funding: The APC was funded by the Open Access Office of the University of Vienna.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Meghan Bohardt very much for her support in the revision process. Tarja Knuutila acknowledges support for this project from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 818772) and Swedish Research Council, grant no. 2018-01353.

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

Abbreviations

The following abbreviations are used in this manuscript:

NOSI Nature of Scientific Inquiry

References

1. Bodner, G.M.; Domin, D.S. Mental Models: The Role of Representations in Problem Solving in Chemistry. *Univ. Chem. Educ.* **2000**, *4*, 24–30.
2. Keiner, L.; Graulich, N. Beyond the Beaker: Students’ Use of a Scaffold to Connect Observations with the Particle Level in the Organic Chemistry Laboratory. *Chem. Educ. Res. Pract.* **2021**, *22*, 146–163. [CrossRef]
3. Lazenby, K.; Stricker, A.; Brandriet, A.; Rupp, C.A.; Mauger-Sonnek, K.; Becker, N.M. Mapping Undergraduate Chemistry Students’ Epistemic Ideas about Models and Modeling. *J. Res. Sci. Teach.* **2020**, *13*, 351. [CrossRef]
4. Schwedler, S.; Kaldewey, M. Linking the Submicroscopic and Symbolic Level in Physical Chemistry: How Voluntary Simulation-Based Learning Activities Foster First-Year University Students’ Conceptual Understanding. *Chem. Educ. Res. Pract.* **2020**, *21*, 1132–1147. [CrossRef]
5. Schwarz, C.V.; Reiser, B.J.; Davis, E.A.; Kenyon, L.; Achér, A.; Fortus, D.; Shwartz, Y.; Hug, B.; Krajcik, J. Developing a Learning Progression for Scientific Modeling: Making Scientific Modeling Accessible and Meaningful for Learners. *J. Res. Sci. Teach.* **2009**, *46*, 632–654. [CrossRef]
6. Cheng, M.F.; Wu, T.Y.; Lin, S.F. Investigating the Relationship Between Views of Scientific Models and Modeling Practice. *Res. Sci. Educ.* **2019**, *51*, 307–323. [CrossRef]
7. Stieff, M.; Scopelitis, S.; Lira, M.E.; Desutter, D. Improving Representational Competence with Concrete Models. *Sci. Educ.* **2016**, *100*, 344–363. [CrossRef]
8. Weisberg, M. Who Is a Modeler? *Br. J. Philos. Sci.* **2007**, *58*, 207–233. [CrossRef]
9. Heidegger, M. *The Question Concerning Technology and Other Essays*; Garland Publishing: New York, NY, USA; London, UK, 1977.
10. Dewey, J. *The Quest for Certainty. A Study of the Relation of Knowledge and Action*; George Allen and Unwin: London, UK, 1929.
11. Nicolaou, C.T.; Constantinou, C.P. Assessment of the Modeling Competence: A Systematic Review and Synthesis of Empirical Research. *Educ. Res. Rev.* **2014**, *13*, 52–73. [CrossRef]
12. Constantinou, C.P.; Nicolaou, C.T.; Papaevripidou, M. A Framework for Modeling-Based Learning, Teaching, and Assessment. In *Towards a Competence-Based View on Models and Modeling in Science Education*; Upmeier zu Belzen, A., Krüger, D., van Driel, J., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 39–58.

13. Rönnebeck, S.; Bernholt, S.; Ropohl, M. Searching for a Common Ground—A Literature Review of Empirical Research on Scientific Inquiry Activities. *Stud. Sci. Educ.* **2016**, *52*, 161–197. [CrossRef]
14. Upmeier zu Belzen, A.; Krüger, D.; van Driel, J. (Eds.) *Towards a Competence-Based View on Models and Modeling in Science Education; Models and Modeling in Science Education*; Springer International Publishing: Cham, Switzerland, 2019; Volume 12. [CrossRef]
15. Lang, V.; Eckert, C.; Perels, F.; Kay, C.W.M.; Seibert, J. A Novel Modelling Process in Chemistry: Merging Biological and Mathematical Perspectives to Develop Modelling Competences. *Educ. Sci.* **2021**, *11*, 611. [CrossRef]
16. Gilbert, J.K.; Justi, R. Models of Modelling. In *Modelling-Based Teaching in Science Education; Models and Modeling in Science Education*; Gilbert, J.K., Justi, R., Eds.; Springer International Publishing: Cham, Switzerland, 2016; Volume 9, pp. 17–40. [CrossRef]
17. Knuuttila, T. Imagination Extended and Embedded: Artifactual versus Fictional Accounts of Models. *Synthese* **2017**, *99*, 56. [CrossRef]
18. Knuuttila, T. Modelling and Representing: An Artefactual Approach to Model-Based Representation. *Stud. Hist. Philos. Sci.* **2011**, *42*, 262–271. [CrossRef]
19. Engelmann, K.; Chinn, C.A.; Osborne, J.; Fischer, F. The Roles of Domain-Specific and Domain-General Knowledge in Scientific Reasoning and Argumentation. An Introduction. In *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*; Routledge: New York, NY, USA, 2018; pp. 1–7.
20. Houkes, W.; Zwart, S.D. Transfer and Templates in Scientific Modelling. *Stud. Hist. Philos. Sci. Part A* **2019**, *77*, 93–100. [CrossRef]
21. Humphreys, P. Knowledge Transfer across Scientific Disciplines. *Stud. Hist. Philos. Sci. Part A* **2019**, *77*, 112–119. [CrossRef]
22. Dunbar, K.N.; Klahr, D. Scientific Thinking and Reasoning. In *The Oxford Handbook of Thinking and Reasoning*; Holyoak, K.J., Morrison, R.G., Eds.; Oxford University Press: Oxford, UK, 2012. [CrossRef]
23. Klahr, D.; Dunbar, K.N. Dual Space Search During Scientific Reasoning. *Cogn. Sci.* **1988**, *12*, 1–48. [CrossRef]
24. Kuhn, D.; Schauble, L. Cross-Domain Development of Scientific Reasoning. *Cogn. Instr.* **1992**, *9*, 285–327. [CrossRef]
25. Löhner, S.; van Joolingen, W.R.; Savelsbergh, E.R.; van Hout-Wolters, B. Students' Reasoning during Modeling in an Inquiry Learning Environment. *Comput. Hum. Behav.* **2005**, *21*, 441–461. [CrossRef]
26. Pedaste, M.; Mäeots, M.; Siiman, L.A.; de Jong, T.; van Riesen, S.A.N.; Kamp, E.T.; Manoli, C.C.; Zacharia, Z.C.; Tsourlidaki, E. Phases of Inquiry-Based Learning: Definitions and the Inquiry Cycle. *Educ. Res. Rev.* **2015**, *14*, 47–61. [CrossRef]
27. Andersen, C.; Garcia-Mila, M. Scientific Reasoning During Inquiry: Teaching for Metacognition. In *Science Education. An International Course Companion*; Taber, K.S., Akpan, B., Eds.; New Directions in Mathematics and Science Education; Sense Publishers: Rotterdam, Poland, 2017; pp. 105–117.
28. Krell, M.; Redman, C.; Mathesius, S.; Krüger, D.; van Driel, J. Assessing Pre-Service Science Teachers' Scientific Reasoning Competencies. *Res. Sci. Educ.* **2020**, *50*, 2305–2329. [CrossRef]
29. Nehring, A.; Stiller, J.; Nowak, K.H.; Upmeier zu Belzen, A.; Tiemann, R. Naturwissenschaftliche Denk- Und Arbeitsweisen Im Chemieunterricht - Eine Modellbasierte Videostudie Zu Lerngelegenheiten Für Den Kompetenzbereich Der Erkenntnisgewinnung. *Z. Für Didakt. Der Naturwissenschaften* **2016**, *22*, 77–96. [CrossRef]
30. Vorholzer, A.; von Aufschnaiter, C.; Kirschner, S. Entwicklung Und Erprobung Eines Tests Zur Erfassung Des Verständnisses Experimenteller Denk- Und Arbeitsweisen. *Z. Für Didakt. Der Naturwissenschaften* **2016**, *22*, 25–41. [CrossRef]
31. Koerber, S.; Osterhaus, C. Individual Differences in Early Scientific Thinking: Assessment, Cognitive Influences, and Their Relevance for Science Learning. *J. Cogn. Dev.* **2019**, *20*, 510–533. [CrossRef]
32. Convertini, J. An Interdisciplinary Approach to Investigate Preschool Children's Implicit Inferential Reasoning in Scientific Activities. *Res. Sci. Educ.* **2021**, *51*, 171–186. [CrossRef]
33. Sodian, B. The Development of Scientific Thinking in Preschool and Elementary School Age. A Conceptual Model. In *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*; Routledge: New York, NY, USA, 2018; pp. 227–250.
34. Hartmann, S.; Upmeier zu Belzen, A.; Krüger, D.; Pant, H.A. Scientific Reasoning in Higher Education. Constructing and Evaluating the Criterion-Related Validity of an Assessment of Preservice Science Teachers' Competencies. *Z. Für Psychol.* **2015**, *223*, 47–53. [CrossRef]
35. Gilbert, J.K. (Ed.) *Visualization in Science Education; Models and Modeling in Science Education*; Springer: Dordrecht, The Netherlands, 2005; Volume 1.
36. Johnson-Laird, P.N. Mental Models and Human Reasoning. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 18243–18250. [CrossRef] [PubMed]
37. Johnson-Laird, P.N.; Byrne, R.M.J.; Schaeken, W. Propositional Reasoning by Model. *Psychol. Rev.* **1992**, *99*, 418–439. [CrossRef]
38. Kind, P.; Osborne, J. Styles of Scientific Reasoning: A Cultural Rational for Science Education? *Sci. Educ.* **2017**, *101*, 8–31. [CrossRef]
39. Samarapungavan, A. Construing Scientific Evidence. The Role of Disciplinary Knowledge in Reasoning with and about Evidence Scientific Practice. In *Scientific Reasoning and Argumentation: The Roles of Domain-Specific and Domain-General Knowledge*; Routledge: New York, NY, USA, 2018; pp. 56–76.
40. Mathesius, S.; Krell, M. Assessing Modeling Competence with Questionnaires. In *Towards a Competence-Based View on Models and Modeling in Science Education; Models and Modeling in Science Education*; Upmeier zu Belzen, A., Krüger, D., van Driel, J., Eds.; Springer International Publishing: Cham, Switzerland, 2019; Volume 12, pp. 117–131. [CrossRef]

41. Göhner, M.; Krell, M. Preservice Science Teachers' Strategies in Scientific Reasoning: The Case of Modeling. *Res. Sci. Educ.* **2020**. [CrossRef]
42. Blömeke, S.; Gustafsson, J.E.; Shavelson, R.J. Beyond Dichotomies. Competence Viewed as a Continuum. *Z. Für Psychol.* **2015**, *223*, 3–13. [CrossRef]
43. Upmeier zu Belzen, A.; van Driel, J.; Krüger, D. Introducing a Framework for Modeling Competence. In *Towards a Competence-Based View on Models and Modeling in Science Education*; Models and Modeling in Science Education; Upmeier zu Belzen, A., Krüger, D., van Driel, J., Eds.; Springer International Publishing: Cham, Switzerland, 2019; Volume 12, pp. 3–19. [CrossRef]
44. Krüger, D.; Kauertz, A.; Upmeier zu Belzen, A. Modelle Und Das Modellieren in Den Naturwissenschaften. In *Theorien in Der Naturwissenschaftsdidaktischen Forschung*; Krüger, D., Parchmann, I., Schecker, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; pp. 141–157. [CrossRef]
45. Upmeier zu Belzen, A.; Krüger, D. Modellkompetenz Im Biologieunterricht. *Z. Für Didakt. Der Naturwissenschaften* **2010**, *16*, 41–57.
46. Giere, R.N. An Agent-Based Conception of Models and Scientific Representation. *Synthese* **2010**, *172*, 269–281. [CrossRef]
47. Gouvea, J.; Passmore, C. 'Models of' versus 'Models For'. *Sci. Educ.* **2017**, *26*, 49–63. [CrossRef]
48. Mahr, B. On the Epistemology of Models. In *Rethinking Epistemology*; Berlin Studies in Knowledge Research; Abel, G., Conant, J., Eds.; De Gruyter: Berlin, Germany; Boston, MA, USA, 2012; Volume 1, pp. 301–352.
49. Stachowiak, H. *Allgemeine Modelltheorie*; Springer: Wien, Austria; New York, NY, USA, 1973.
50. Chamizo, J.A. The Role of Instruments in Three Chemical' Revolutions. *Sci. Educ.* **2014**, *23*, 955–982. [CrossRef]
51. Espinet, M.; Izquierdo, M.; Bonil, J.; Ramos de Robles, S.L. The Role of Language in Modeling the Natural World: Perspectives in Science Education. In *Second International Handbook of Science Education*; Number 24 in Springer International Handbooks of Education; Springer: Dordrecht, The Netherlands; Heidelberg, Germany; London, UK; New York, NY, USA, 2012; pp. 1385–1403.
52. Justi, R.S.; Gilbert, J.K. Modelling, Teachers' Views on the Nature of Modelling, and Implications for the Education of Modellers. *Int. J. Sci. Educ.* **2002**, *24*, 369–387. [CrossRef]
53. van der Valk, T.; van Driel, J.H.; de Vos, W. Common Characteristics of Models in Present-Day Scientific Practice. *Res. Sci. Educ.* **2007**, *37*, 469–488. [CrossRef]
54. Matthews, M.R. Models in Science and in Science Education: An Introduction. *Sci. Educ.* **2007**, *16*, 647–652. [CrossRef]
55. Tang, K.S. The Use of Epistemic Tools to Facilitate Epistemic Cognition & Metacognition in Developing Scientific Explanation. *Cogn. Instr.* **2020**, *38*, 474–502. [CrossRef]
56. Thomas, G.P. 'Triangulation:' An Expression for Stimulating Metacognitive Reflection Regarding the Use of 'Triplet' Representations for Chemistry Learning. *Chem. Educ. Res. Pract.* **2017**, *18*, 533–548. [CrossRef]
57. Nersessian, N.J. Model-Based Reasoning in Conceptual Change. In *Model-Based Reasoning in Scientific Discovery*; Magnani, L., Nersessian, N.J., Thagard, P., Eds.; Springer: New York, NY, USA, 1999; pp. 5–22.
58. Carey, S.; Evans, R.; Honda, M.; Jay, E.; Unger, C. 'An Experiment Is When You Try It and See If It Works': A Study of Grade 7 Students' Understanding of the Construction of Scientific Knowledge. *Int. J. Sci. Educ.* **1989**, *11*, 514–529. [CrossRef]
59. Passmore, C.; Gouvea, J.; Giere, R.N. Models in Science and in Learning Science: Focusing Scientific Practice on Sense-making. In *International Handbook of Research in History, Philosophy and Science Teaching*; Matthews, M.R., Ed.; Springer: Dordrecht, The Netherlands, 2014; pp. 1171–1202.
60. Devitt, M. Scientific Realism. In *Truth and Realism*; Greenough, P., Lynch, M.P., Eds.; Oxford University Press: Oxford, UK, 2006; pp. 100–124. [CrossRef]
61. Reith, M.; Nehring, A. Scientific Reasoning and Views on the Nature of Scientific Inquiry: Testing a New Framework to Understand and Model Epistemic Cognition in Science. *Int. J. Sci. Educ.* **2020**, *42*, 2716–2741. [CrossRef]
62. Caspari, I.; Weinrich, M.L.; Sevan, H.; Graulich, N. This Mechanistic Step Is "Productive": Organic Chemistry Students' Backward-Oriented Reasoning. *Chem. Educ. Res. Pract.* **2018**, *19*, 42–59. [CrossRef]
63. Caspari, I.; Kranz, D.; Graulich, N. Resolving the Complexity of Organic Chemistry Students' Reasoning through the Lens of a Mechanistic Framework. *Chem. Educ. Res. Pract.* **2018**, *19*, 1117–1141. [CrossRef]
64. Louca, L.T.; Zacharia, Z.C. Examining Learning Through Modeling in K-6 Science Education. *J. Sci. Educ. Technol.* **2015**, *24*, 192–215. [CrossRef]
65. Oliva, J.M.; del Mar Aragón, M.; Cuesta, J. The Competence of Modelling in Learning Chemical Change. *Int. J. Sci. Math. Educ.* **2015**, *13*, 751–791. [CrossRef]
66. Fox Keller, E. Models of and Models for: Theory and Practice in Contemporary Biology. *Philos. Sci.* **2000**, *67*, 72–86. [CrossRef]
67. Suárez, M. An Inferential Conception of Scientific Representation. *Philos. Sci.* **2004**, *71*, 767–779. [CrossRef]
68. Suárez, M. Scientific Representation. *Philos. Compass* **2010**, *5*, 91–101. [CrossRef]
69. Knuuttila, T. Epistemic Artifacts and the Modal Dimension of Modeling. *Eur. J. Philos. Sci.* **2021**, *11*, 65. [CrossRef]
70. Taber, K.S. Models and Modelling in Science and Science Education. In *Science Education. An International Course Companion*; Number 31 in New Directions in Mathematics and Science Education; Springer: Dordrecht, The Netherlands, 2017; pp. 263–278.
71. Bueno, O.; Colyvan, M. An Inferential Conception of the Application of Mathematics. *Noûs* **2011**, *45*, 345–374. [CrossRef]
72. Chakravartty, A. Informational versus Functional Theories of Scientific Representation. *Synthese* **2010**, *172*, 197–213. [CrossRef]
73. Suárez, M. The Modelling Attitude and Its Roots in 19th Century Science. 2014. Available online: <https://scholarworks.iu.edu/dspace/handle/2022/26193> (accessed on 11 April 2022).
74. Boltzmann, L. Model. In *Encyclopedia Britannica*, 11th ed.; Cambridge University Press: London, UK, 1902; pp. 211–220.

75. Wartofsky, M.W. The Model Muddle: Proposals for an Immodest Realism. In *Models. Representation and the Scientific Understanding*; Boston Studies in the Philosophy of Science; Springer: Dordrecht, The Netherlands, 1979; Volume 48, pp. 1–11.
76. Suárez, M. Scientific Representation: Against Similarity and Isomorphism. *Int. Stud. Philos. Sci.* **2003**, *17*, 225–244. [CrossRef]
77. Frigg, R. Scientific Representation and the Semantic View of Theories. *Theoria* **2006**, *21*, 49–65.
78. Bartels, A. Defending the Structural Concept of Representation. *Theoria* **2006**, *21*, 7–19.
79. Bueno, O.; French, S. How Theories Represent. *Br. J. Philos. Sci.* **2011**, *62*, 857–894. [CrossRef]
80. Giere, R.N. How Models Are Used to Represent Reality. *Philos. Sci.* **2004**, *71*, 742–752. [CrossRef]
81. Morrison, M.; Morgan, M.S. Models as Mediating Instruments. In *Models as Mediators. Perspectives on Natural and Social Science*; Morgan, M.S., Morrison, M., Eds.; Cambridge University Press: Cambridge, UK; New York, NY, USA; Melbourne, Australia; Madrid, Spain; Cape Town, South Africa; Singapore; Sao Paulo, Brazil, 1999; pp. 10–37.
82. Mäki, U. Contested Modeling: The Case of Economics. In *Models, Simulations, and the Reduction of Complexity*; Abhandlungen Der Akademie Der Wissenschaften in Hamburg; De Gruyter: Berlin, Germany; New York, NY, USA, 2013; Volume 4, pp. 87–106.
83. Kingsland, S.E. *Modeling Nature: Episodes in the History of Population Ecology*, 2nd ed.; Science and Its Conceptual Foundations; University of Chicago Press: Chicago, IL, USA, 1995.
84. Weisberg, M. The Anatomy of Models. In *Simulation and Similarity. Using Models to Understand the World*; Oxford University Press: New York, NY, USA, 2013; pp. 24–45.
85. Tversky, A. Features of Similarity. *Psychol. Rev.* **1977**, *84*, 327–352. [CrossRef]
86. Preston, B. Artifact. In *The Stanford Encyclopedia of Philosophy*; Metaphysic Research Lab: Stanford, CA, USA, 2020. Available online: <https://plato.stanford.edu/archives/fall2020/entries/artifact/> (accessed on 11 April 2022).
87. Knuuttila, T. Models, Fictions and Artifacts. In *Language and Scientific Research*; Springer International Publishing: Cham, Switzerland, 2021; pp. 199–220.
88. Volterra, V. Variations and Fluctuations of the Number of Individuals in Animal Species Living Together. *ICES J. Mar. Sci.* **1928**, *3*, 3–51. [CrossRef]
89. Whittaker, E.T. Vito Volterra, 1860—1940. *Obit. Not. Fellows R. Soc.* **1941**, *3*, 691–729. [CrossRef]
90. May, R.M. Biological Populations with Nonoverlapping Generations: Stable Points, Stable Cycles, and Chaos. *Science* **1974**, *186*, 645–647. [CrossRef]
91. Knuuttila, T.; Loettgers, A. Model Templates within and between Disciplines: From Magnets to Gases—and Socio-Economic Systems. *Eur. J. Philos. Sci.* **2016**, *6*, 377–400. [CrossRef]
92. Strevens, M. *Depth: An Account of Scientific Explanation*; Harvard University Press: Cambridge, MA, USA; London, UK, 2008.
93. Mäki, U. On the Method of Isolation in Economics. *Pozn. Stud. Philos. Sci. Humanit.* **1992**, *26*, 19–54.
94. Carrillo, N.; Knuuttila, T. An Artifactual Perspective on Idealization: Constant Capacitance and the Hodgkin and Huxley Model. In *Models and Idealizations in Science. Artifactual and Fictional Approaches*; Logic, Epistemology, and the Unity of Science; Cassini, A.; Redmond, J., Eds.; Springer International Publishing: Cham, Switzerland, 2021; Volume 50, pp. 51–70. [CrossRef]
95. Rice, C. Idealized Models, Holistic Distortions, and Universality. *Synthese* **2018**, *195*, 2795–2819. [CrossRef]
96. Knuuttila, T.; Loettgers, A. Biological Control Various Materialized: Modeling, Experimentation and Exploration in Multiple Media. *Perspect. Sci.* **2021**, *29*, 468–492. [CrossRef]
97. Kress, G.R.; van Leeuwen, T. *Multimodal Discourse*; Bloomsbury Academic: London, UK, 2001.
98. Landy, D.; Allen, C.; Zednik, C. A Perceptual Account of Symbolic Reasoning. *Front. Psychol.* **2014**, *5*, 275. [CrossRef] [PubMed]
99. Johansen, M.W.; Misfeldt, M. Material Representations in Mathematical Research Practice. *Synthese* **2020**, *197*, 3721–3741. [CrossRef]
100. Morgan, M.S.; Boumans, M.J. Secrets Hidden by Two-Dimensionality: The Economy as Hydraulic Machine. In *Models: The Third Dimension of Science*; Writing Science, Stanford University Press: Stanford, CA, USA, 2004.
101. Gelfert, A. Exploratory Uses of Scientific Models. In *How to Do Science with Models*; SpringerBriefs in Philosophy; Springer International Publishing: Cham, Switzerland, 2016; pp. 71–99. [CrossRef]
102. Godfrey-Smith, P. The Strategy of Model-Based Science. *Biol. Philos.* **2006**, *21*, 725–740. [CrossRef]
103. Boumans, M.J. Built-in Justification. In *Models as Mediators. Perspectives on Natural and Social Science*; Cambridge University Press: Cambridge, UK, 1999; pp. 66–96.
104. Sjöström, J. Towards Bildung-Oriented Chemistry Education. *Sci. Educ.* **2013**, *22*, 1873–1890. [CrossRef]
105. Reid, N. Johnstone’s Triangle: Why Chemistry Is Difficult. In *The Johnstone Triangle: The Key to Understanding Chemistry*; Royal Society of Chemistry: Cambridge, UK, 2021; pp. 48–71. [CrossRef]
106. Hoyningen-Huene, P. The Systematicity of Science Unfolded. In *Systematicity: The Nature of Science*; Oxford Studies in Philosophy of Science; Oxford University Press: New York, NY, USA, 2013.
107. Flores-Camacho, F.; Gallegos-Cázares, L.; Garritz, A.; García-Franco, A. Incommensurability and Multiple Models: Representations of the Structure of Matter in Undergraduate Chemistry Students. *Sci. Educ.* **2007**, *16*, 775–800. [CrossRef]
108. Gobert, J.D.; O’Dwyer, L.; Horwitz, P.; Buckley, B.C.; Levy, S.T.; Wilensky, U. Examining the Relationship Between Students’ Understanding of the Nature of Models and Conceptual Learning in Biology, Physics, and Chemistry. *Int. J. Sci. Educ.* **2011**, *33*, 653–684. [CrossRef]
109. Krell, M.; Reinisch, B.; Krüger, D. Analyzing Students’ Understanding of Models and Modeling Referring to the Disciplines Biology, Chemistry, and Physics. *Res. Sci. Educ.* **2015**, *45*, 367–393. [CrossRef]

110. Treagust, D.F.; Chittleborough, G.; Mamiala, T.L. Students' Understanding of the Role of Scientific Models in Learning Science. *Int. J. Sci. Educ.* **2002**, *24*, 357–368. [CrossRef]
111. Daniel, K.L.; Bucklin, C.J.; Austin Leone, E.; Idema, J. Towards a Definition of Representational Competence. In *Towards a Framework for Representational Competence in Science Education; Models and Modeling in Science Education*; Daniel, K.L., Ed.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11, pp. 3–11. [CrossRef]
112. Stieff, M.; DeSutter, D. Sketching, Not Representational Competence, Predicts Improved Science Learning. *J. Res. Sci. Teach.* **2021**, *58*, 128–156. [CrossRef]

Article

Modeling as Scientific Reasoning—The Role of Abductive Reasoning for Modeling Competence

Annette Upmeier zu Belzen ^{1,*} , Paul Engelschalt ¹ and Dirk Krüger ² ¹ Biology Education, Humboldt-Universität zu Berlin, 10099 Berlin, Germany; paul.engelschalt@hu-berlin.de² Biology Education, Freie Universität Berlin, 14195 Berlin, Germany; dirk.krueger@fu-berlin.de

* Correspondence: annette.upmeier@biologie.hu-berlin.de

Abstract: While the hypothetico-deductive approach, which includes inductive and deductive reasoning, is largely recognized in scientific reasoning, there is not much focus on abductive reasoning. Abductive reasoning describes the theory-based attempt of explaining a phenomenon by a cause. By integrating abductive reasoning into a framework for modeling competence, we strengthen the idea of modeling being a key practice of science. The framework for modeling competence theoretically describes competence levels structuring the modeling process into model construction and model application. The aim of this theoretical paper is to extend the framework for modeling competence by including abductive reasoning, with impact on the whole modeling process. Abductive reasoning can be understood as knowledge expanding in the process of model construction. In combination with deductive reasoning in model application, such inferences might enrich modeling processes. Abductive reasoning to explain a phenomenon from the best fitting guess is important for model construction and may foster the deduction of hypotheses from the model and further testing them empirically. Recent studies and examples of learners' performance in modeling processes support abductive reasoning being a part of modeling competence within scientific reasoning. The extended framework can be used for teaching and learning to foster scientific reasoning competences within modeling processes.

Keywords: scientific reasoning; abductive reasoning; models; modeling; model construction; model application; modeling competence

Citation: Upmeier zu Belzen, A.; Engelschalt, P.; Krüger, D. Modeling as Scientific Reasoning—The Role of Abductive Reasoning for Modeling Competence. *Educ. Sci.* **2021**, *11*, 495. <https://doi.org/10.3390/educsci11090495>

Academic Editor: Gavin T. L. Brown

Received: 6 August 2021

Accepted: 29 August 2021

Published: 3 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Theoretical abduction is “the process of reasoning in which explanatory hypotheses are formed and evaluated” [1] (p. 220).

This process of reasoning addresses modeling. Hence, the concept about the phenomenon is a model [2] that develops while seeking for explanations [3]. Thus, scientific reasoning, in terms of searching for explanations to obtain insight into a phenomenon, is related to the construction of models. The derivation of hypotheses from these models and their application in empirical investigations allows the evaluation of the phenomenon [1]. As such, modeling is a prominent style of scientific reasoning that also is understood as a skill that needs to be practiced [4] and is related to competences [5] (p. 43). Thus, the framework for modeling competence was developed [6,7], respecting particularly procedural and epistemic perspectives of reasoning [8].

A model serves as a representation for communicating scientific knowledge or as a research tool for testing hypotheses about a phenomenon [9]. A model used for teaching and learning content knowledge serves as a medium for communication and meets the purpose of describing and explaining current scientific knowledge. Therefore, when using a model as medium, the focus is on model construction to represent a phenomenon accurately [7,9–11]. In contrast, a model used as a research tool is constructed for the purpose of deriving hypotheses about scientific phenomena. Hence, the focus is on model

application in research contexts to gain new insights into unknown phenomena [11]. In both communication and research contexts, model construction and model application are central parts of intertwined modeling processes: a model is constructed starting from a theoretical background and from abductive or inductive reasoning [12,13], both forms of logical inferences [14]. Deductive reasoning as the third logical inference [14] is practiced in model application, which starts with deriving hypotheses deductively from the model, usually followed by empirical testing [4].

In biology education, the framework for modeling competence (FMC) [6] has been developed and empirically validated [15,16]. The FMC structures modeling competence into aspects and levels [17] and addresses at the same time the perspectives of model construction and model application. Theoretical considerations and empirical findings [3,18] revealed the need for including another level to the FMC regarding reasoning processes in model construction with hypothesized impact for model application [7]. This extension was realized by integrating the knowledge-expanding function of explaining a phenomenon in the process of model construction, which is abductive reasoning [3,13,19]. In the initial FMC, explaining was considered as an intermediate level representing communicative functions. However, this approach did not cover the idea of developing a model by explaining a phenomenon with causes from past experiences and information [3,13,19], meaning a phenomenon is explained as best as possible through abductive reasoning [13]. Thus, the term “explanation” [3,18] describes two different practices needing to be separated: explanation in order “to make clear” for communication purposes and explanation in order “to justify” as an epistemic function. This differentiation of explaining is now integrated into the presented FMC. The process of abductive reasoning in model construction may initiate, because of its uncertainty, deductively derived hypotheses in model application and thus promote empirical investigations.

In this article, we argue that abductive and deductive reasoning are related parts within scientific reasoning regarding model construction and model application. The theoretical considerations of abductive reasoning in modeling are supported by empirical work in mathematics [20] and geography [12,21,22]. Additionally, we give some insight into learners’ performance in modeling processes which support abductive reasoning being a part of modeling competence within scientific reasoning.

2. Logical Reasoning

Three forms of logical reasoning are involved in scientific reasoning and inquiry. They are summarized briefly by Peirce: “The division of all inference into Abduction, Deduction, and Induction may almost be said to be the Key of Logic” [14] (CP 2.98). In this context, abduction is about generating a cause as the best explanation for an observed phenomenon based on existing rules or theoretical knowledge (“inference to the best explanation” [23], “educated guess” [12]). This kind of reasoning is knowledge expanding, leads to creative ideas, and thus forms new theoretical inferences [24]. In contrast, inductive reasoning derives a general rule from repeated observations of a phenomenon. This inference is knowledge expanding but does not provide any principally new ideas [14]. In deductive reasoning, a general rule as theoretical basis and a cause are used to predict a result of a certain case. If the rule is true, each individual case will fit to this rule. Thus, deductive reasoning is truth preserving and logically flawless. However, as in the case of inductive reasoning, it does not generate principally “new ideas” [24]. The relationship among the three forms of logical reasoning is summarized by Peirce: “Deduction proves that something must be; Induction shows that something actually is operative; Abduction merely suggests that something may be” [14] (CP 5.171).

3. Theory of Abductive Reasoning

An established theory of abductive reasoning from cognitive psychology describes seven components of abductive reasoning [13,25]. This theory describes a continuous, implicit process with different steps that do not have to be run through in a strict order [26].

This process can lead to a consistent type of explanation free from redundancies [13]. Ideally, the process of abductive reasoning begins with the perception of a phenomenon, for which the step of *data collection* takes place in an exploratory or theory-based manner. Subsequently, these data are incorporated into an existing mental model leading to a preliminary *comprehension*. It is *checked* whether the new data contradict the previous model or remain un-understandable. These thoughts lead to the step of *resolving anomaly*. If this occurs, new *data* will be *collected*. If there are several possible explanations, alternative potentially plausible explanations will be *refined*. Due to this, it is necessary to *discriminate* by selecting one potentially plausible explanation. In the step of *checking* for consistency, both likely and unlikely explanations are included. This process of decision making may lead to the *collection* of new *data*. If *checking* for consistency is not successful, other potentially plausible explanations will be *discriminated*. Although model *testing* in the theory of Johnson and Krems is about eliminating this uncertainty about improbable explanations [13], this step can be extended to an abductively developed model. When it comes to application of this model, hypotheses are derived deductively to be *tested* (“*abductive model evaluation*”) [1,8].

4. Models and Modeling

4.1. Concept of a Model

“In model-based views, models are considered subsets of scientific theories—more comprehensive systems of explanations—which are created with various semiotic resources and provide semantically rich information for scientific reasoning and problem solving” [27] (p. 1110).

The term model has so many meanings that attempts merging all meanings into one definition are methodologically useless [28]. Hence, there is no unified definition of what a model in science and science education is [29,30], nor is there a unifying modeling theory [31]. Following Mittelstraß, models are replicas of a real or imaginary object with the aim of learning something about it or learning something with it [32]. This refers to both the representational function (learning something about it) of models for the purpose of communication and to the research tool function (learning something with it) to test new ideas for the purpose to generate new knowledge.

Due to the multiformity of models and since anything can become a model that is conceived of something as a model by an agent for some purpose and time [4,33–35], general properties that characterize models ontologically as special objects are absent [32,34]. Other approaches distance from an ontological perspective on models and try to conceptualize models from an epistemic point of view [30,34,36]. In this case, something becomes a model when it is used [4], developed [31], or conceived as such [34]. In his concept of model-being, Mahr suggests that an agent judges something to be a model for a specific period of time and for a specific purpose [34]. Furthermore, the distinction between the imagined mental model and the externalized model object is relevant to Mahr’s conceptualization of models [26]. In this context, the model object is described as the representation of a mental model in the broadest sense, reaching from verbal analogies to graphical representations.

4.2. Concept of Model-Being

The model and the model object each stand in two relations to something: in the perspective of construction, the model stands in relation to something of which it is a model. In the perspective of application, it stands in relation to something for which it is used for as a model [9,28,34] (Figure 1). These two relationships are constitutive and inherent aspects of model-being [28,36,37].

Mahr’s concept of model-being has separated inherent properties permanently associated with the model object [34]. A model can be used by an agent as a model *of* something and *for* something in any given time.

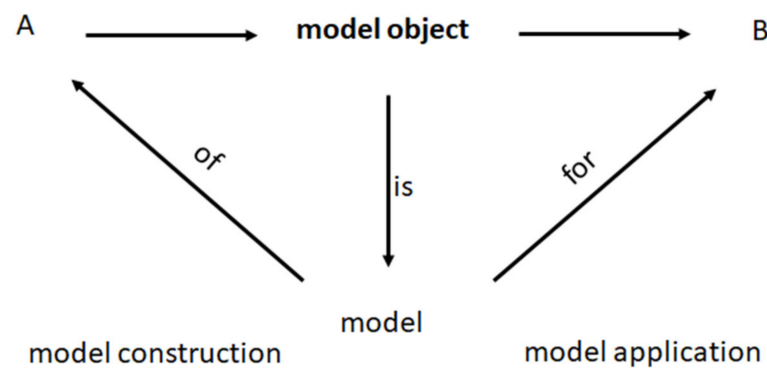


Figure 1. Mahr's concept of model-being [7] (adapted). A is a phenomenon, B is a purpose of an application.

Gouvea and Passmore also differentiated into the perspectives models *for* something (as tools for research) and models *of* something (as representation of actual knowledge) [9]. In contrast to Mahr's concept of model-being, their categorization of these perspectives are not constitutive aspects of a model in terms of a theoretical understanding of model-being. Gouvea and Passmore rather argue from a heuristic perspective to help teachers and supporting students. In accordance with this perspective, Gilbert and Justi suggest conceiving models as substitutes [38] or to describe models as epistemic tools [31] being used by agents [30].

Giere described the agent as the person making decisions about both the focus of the similarities (intent) and the goal of that focus (purpose) [36]. Mahr also consistently integrates an agent in his concept of model-being [39]. He distinguishes between the mental model, which is modeled by the agent, and the model object as the externalized representation of the agent's mental model.

4.3. Modeling Process

The process of modeling lacks a general procedural description and definition of certain rules [40]. This is because experiences, ideas, and theories of the modeling agent influence the process and hence creative, innovative, and subjective considerations are involved [12,41]. Nevertheless, recurring elements can be identified in modeling, which ideally follow a hypothetico-deductive research logic [42,43]. In the following, the process scheme of modeling described by Krell and colleagues [15,16] stands as the basis for the integration of abductive reasoning.

In the scheme, the modeling process begins with the perception of a phenomenon, most frequently undertaken by observation (Figure 2) [44]. Observation in this case means exploring the phenomenon as a whole and without explicit assumptions [44]. These observations might lead to the formulation of hypotheses about potential relations between variables, which means that conceivable theories are generated. These hypotheses can arise through inductive reasoning from a generalized model. They are checked for consistency with other theories within model construction (Figure 2). Alternatively, abductive reasoning explains the phenomenon [13], for example with the help of analogies and is also checked for consistency (Figure 2). In case consistency is missing, the phenomenon is further explored by additional observations. If inductive or abductive inferences lead to plausible models, model construction temporarily ends. Model application begins with the deduction of hypotheses about how the model's relationships will behave under certain conditions (Figure 2). Depending on the type of hypotheses, this leads to different methodological implementations and thus into corresponding inquiry methods [6,45]. While difference hypotheses are descriptive and lead to the comparison of structures, groups, or systems, causal hypotheses are investigated through controlled experimentation and correlation hypotheses through observation (Figure 2) [44]. The analysis of data from empirical investigations lead to support or falsification of hypotheses (Figure 2). If sources

of interference in data collection are excluded as a reason for the lack of fit between the hypotheses derived from the model and the phenomenon under investigation, the model, the model object, and the concept about the modeled phenomenon have to be revised. In this process, exploration of the phenomenon restarts, which means that the process of model construction and application of a modified model begins anew (Figure 2) [4]. By initiating cognitive processes this way, models become flexible intellectual tools for scientific knowledge acquisition (epistemic tools) [46,47]. This function goes beyond presenting a model of something in a medial perspective as a means for communication.

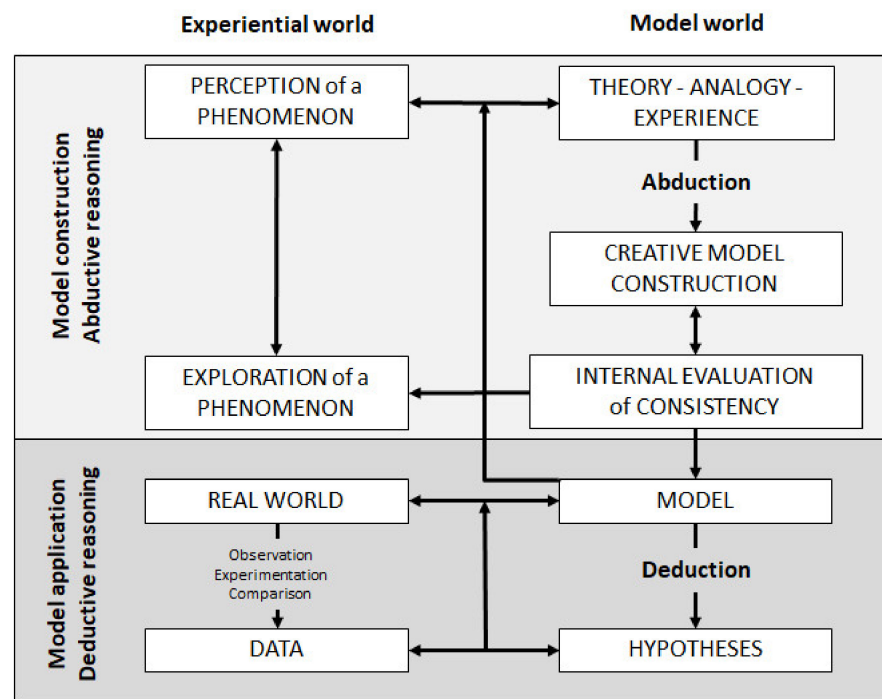


Figure 2. Abductive reasoning in model construction and deductive reasoning in model application.

5. Framework for Modeling Competence

The initial FMC [10] structures modeling competence in five aspects and three levels. The aspects were built on the basis of studies from science education research [48–50]: *nature of models*, *multiple models*, *purpose of modeling*, *testing models*, and *changing models* [6,10,15] (Figure 3). In the case of *nature of models*, the focus is on the similarity between the model and the phenomenon. The aspect *alternative models* addresses the question whether there can exist several models for a phenomenon. The *purpose of modeling* is guiding the modeling process for communication or as a research tool. Considering the purpose, when *testing models* and *changing models* from a medial perspective, it is about optimizing the model in context of already known details. In the research tool perspective, *testing models* and *changing models* starts from hypotheses and is led by results from corresponding empirical investigations.

The three competence levels were based on Mahr's conceptualization of model-being and integrate perspectives on modeling focusing on the model object (level I), model construction (level II), and model application (level IIIb, Figure 3) [10]. The extended FMC integrates abductive reasoning as a further level (Figure 3, level IIIa) [7]. This new level differs from an understanding-generating explanation of common knowledge with models of something determining level II (Figure 3). In contrast, the knowledge-expanding function of explaining described in level IIIa is based on abductive reasoning. Abductive reasoning in model construction, like deductive reasoning in model application, involves theoretical or creative considerations. By treating level IIIa as part of level III, it is intended to clarify that model construction by abductive reasoning is scientifically demanding [13].

It may precede deductive reasoning in the sense of the hypothetico-deductive path of knowledge acquisition [43,51] (Figure 3).

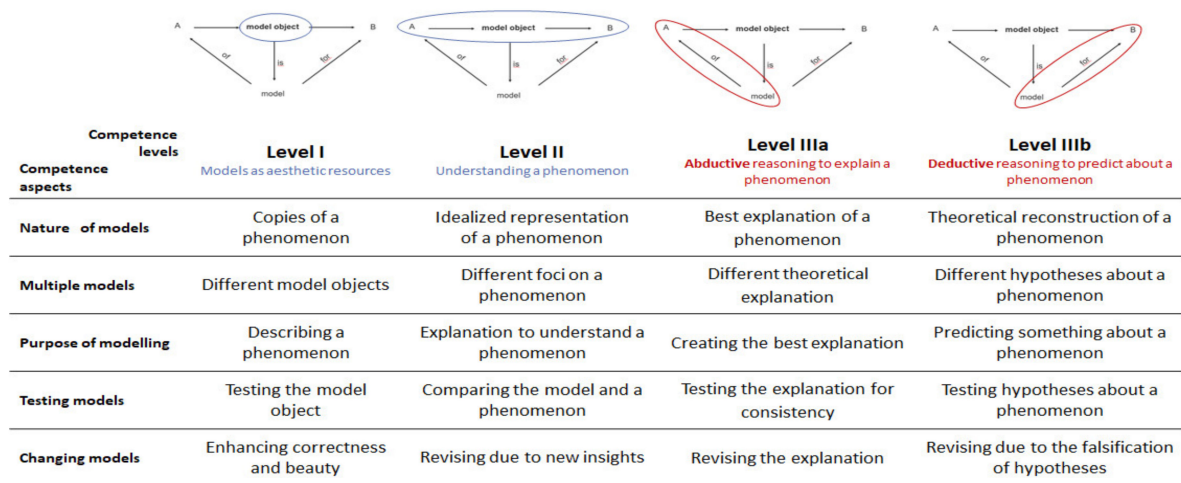


Figure 3. Framework for modeling competence [6,7] consisting of aspects and levels.

The inclusion of a general theory of abductive reasoning in modeling [26,52], in which a model is constructed sequentially in a complex and creative process of understanding, leads to a definition of the cognitive facet of modeling competence: modeling competence comprises the abilities to initiate a theory-guided cognitive process in the creative construction of models, to gain purpose-related knowledge in the application of models, to judge about models with reference to their purpose, and to reflect on the modeling process in terms of scientific reasoning [7].

In the extended FMC (Figure 3), thinking about models and modeling that is assigned to levels I and II means to understand models and modeling as representations to achieve educational goals, which is the medial perspective [6,7,10,11]. The focus is on accuracy in model construction for communication, teaching, and learning of content knowledge. In more detail, level I deals with the ability to assess the model object from an aesthetic point of view or regarding its technical functionality without putting the phenomenon in relation to the model object, except in its capacity as a copy or for the purpose of illustration. Level II entails the ability to assess the process of model construction for understanding the represented phenomenon. The model object is a more or less accurate representation of something already known in the natural sciences.

Descriptions in level IIIa and IIIb indicate an understanding of modeling in the context of scientific investigations, which means the ability to assess models in their construction and application as research tools, which is a methodological perspective [6,7,10,11]. Level IIIa describes the ability to construct a model that provides the best plausible explanation for unknown phenomena which is free of contradictions to previous theories and explorations. Modeling is thus already a theoretical or creative process in model construction, which, associated with uncertainty, represents knowledge about a phenomenon, and can offer new possibilities for explanation. Level IIIb describes the ability to apply a model as a tool for investigating a phenomenon within scientific reasoning to empirically test its validity in the hypothetico-deductive approach; the model object as a model for something leads to processing new, thus far unexplained scientific questions.

The competence descriptions with regard to aspects and levels draw on theoretical elaborations [10,11] and, with regard to the initial FMC, extensive empirical work, which allows the use of them for assessing and promoting modeling competence for scientific reasoning [6,15,53]. However, level IIIa of the FMC with inclusion of abductive reasoning still needs to be empirically investigated.

6. State of Research

There are several approaches from different disciplines of science education connecting abductive reasoning with modeling [12,22,54]. For geoscience, Oh established a close connection between abduction and modeling (modeling-based abductive reasoning) [21,22] relating to research by Clement (addressing the solution of physical problems through abductive reasoning in modeling) [12]. Furthermore, Park and Lee point to the central role of abductive reasoning in mathematical modeling [20]. These studies rather focus on the role of abductive reasoning for constructing technically appropriate models in terms of content knowledge than on methodological (procedural and epistemic) knowledge as part of scientific reasoning competencies.

Our work aims to obtain insight into abductive reasoning within modeling processes in biological contexts. Based on Sturm [55], the reddened face phenomenon was used to obtain insight into abductive reasoning with regard to the FMC's competence descriptions [6]. Regarding this, 32 pre-service biology teachers created concept maps to solve the problem of why a fictitious person, whom they cannot talk to, has a reddened face. The participants generated abductive explanations and strategies for testing these explanations (Figure 4). A total of 159 explanations were summarized into 39 types of explanations for the reddened face. It turns out that the reddened face scenario promotes students to select different explanations by abductive reasoning. Most of the 39 given explanations were further condensed into six superior explanation types "Emotion", "Activity", "Disease", "Environment", "Blood Circulation", and "Individual Disposition" (Figure 4).

Explanation types	Single explanations	N explanations	
		N explanations	N tested explanations
Emotion	<i>shame, stress, fury, anxiety, love, nervousness, excitement, anger, sadness, uncomfortable situation, emotional status, dispute, shyness, remorse, bad mood, worries</i>	62	19
Activity	<i>physical exertion, sport</i>	26	14
Disease	<i>fever, sickness, allergic reaction, intolerance, Lupus erythematoses, scarlet</i>	23	6
Environment	<i>temperature, sun, environmental factors, humidity</i>	21	13
Blood Circulation	<i>high blood circulation, high blood pressure</i>	11	1
Individual Disposition	<i>skin color, normal look, genetic reason, overweight</i>	6	1
Other Explanation	<i>alcohol, color in the face, spicy food, cosmetic treatment</i>	10	3

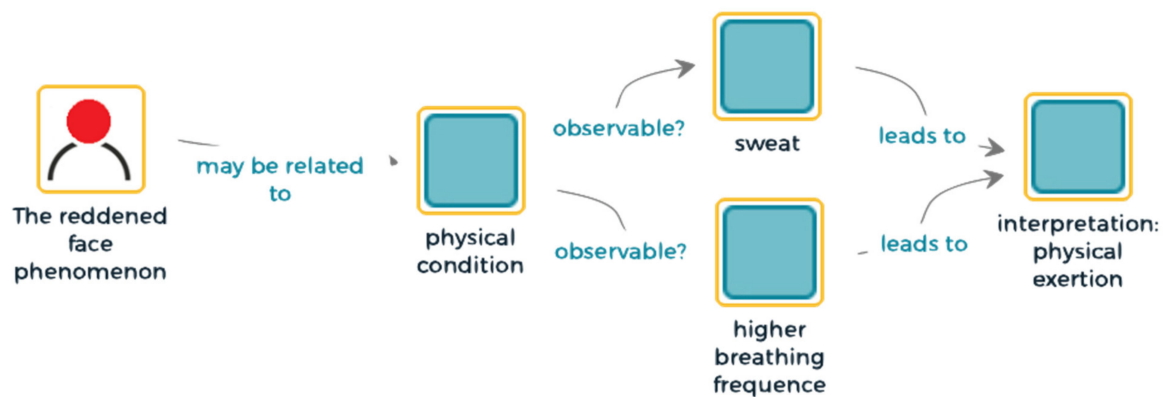
Figure 4. Frequencies of explanations ($N = 159$) and of tested explanations ($N = 57$) per explanation type.

In total, for 57 of 159 explanations further considerations for testing were provided. Hence, most explanations were not linked with ideas on how to test them. Explanations such as "Activity" or "Environment" have been connected to possible test strategies most frequently. In everyday life, explanations regarding "Emotion" can be tested easily through verbal communication. As this was not possible, this may explain why the participants tested explanations for "Activity" or "Environment" more frequently than for "Emotion".

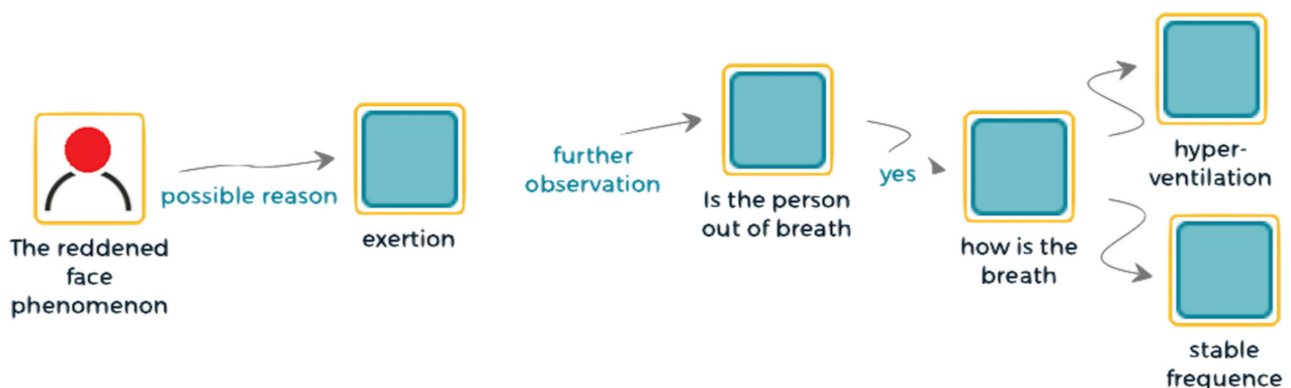
The reddened face phenomenon emphasizes that a complex and creative process of abductive reasoning [13] is relevant in model construction [22] and is successful as soon as experience and analogies allow for abductive reasoning. Studies showed that for modeling the inner mechanisms in black-box scenarios [47–49], abductive reasoning

can be a successful strategy when a theoretical background is available, or creativity is involved when interpreting data. This can lead to repeated switching between abductive and deductive reasoning. If the development of explanations for the inner mechanisms of the black box is not satisfactory [56], this is because theoretical knowledge is not available, analogies are not found, or creative solutions are lacking [57]. Unsurprisingly, because a corresponding model is missing, this leads to neither model application nor deductive reasoning [56,58].

Students' solutions for the reddened face phenomenon were structured into three different groups. In the first group, possible explanations were simply guessed without any testing strategy ($n = 12$). This result does not fit to our expectation that abductive explanations in model construction foster the switch into deductive testing in model application on its own. On the other hand, this result may be related to the fact that there were no possibilities to interact with the fictitious person nor the phenomenon itself. Thus, there was no feedback or interactive offer to test explanations. Nonetheless, most concept maps ($n = 20$) provided indicators aiming to test abducted explanations. Among these, two different strategies were identified. The first strategy in the sense of *abductive testing* ($n = 7$, Figure 5a) is characterized by the derivation of an explanation from additional speculatively observed indicators (test; cf. [13]). In contrast, the second strategy in the sense of *deductive testing* ($n = 13$, Figure 5b) is characterized by indicators for further observations being derived from a possible explanation. The strategy of abductive testing refers to the theory of abductive reasoning [13] by collecting further information beforehand within observations (Exploration of the phenomenon, Figure 2), thus in model construction. By switching to model application, applying the strategy of deductive testing of abducted explanations, students indicate strategies of deductive reasoning. This result supports the idea that abductive reasoning in model construction fosters strategies for deductive reasoning in model application.



(a) Abductive testing



(b) Deductive testing

Figure 5. Excerpts of students' concept maps illustrating the strategies of abductive testing (a) and deductive testing (b).

7. Outlook

The focus on abductive reasoning within modeling processes is rather new [20,21] and led to the extension of the FMC for the field of biology in the natural sciences, thus providing a theoretical basis for the investigation of scientific reasoning in this modeling perspective. This innovation can be referred to as “abductive turn” [59], leading to broader foundations of scientific reasoning in terms of paths of knowledge acquisition in science education [6,51]. Explicating the role of induction when encountering a phenomenon, the role of abduction in model construction, and the role of deduction in model application supports Lehrer and Schauble’s suggestion to consider modeling as the “signature practice of science” [60]. In this way, the prominent position of induction and deduction within the hypothetico-deductive approach might be expanded by integrating abductive reasoning in the classroom, with implications for Nature of Science perspectives [8,61]. It is necessary to further reflect on the role of abduction for gaining new knowledge and to answer the question whether abductive reasoning is underrepresented compared to induction and deduction in the hypothetico-deductive approach [62,63]. In other words, the focus on deductive inference may fall too short [64], and abduction should be implemented in school curricula as an important part of scientific reasoning.

Taking the reported rare empirical insight about the role of abductive reasoning for modeling into account, it becomes clear that scientific reasoning in modeling leads to considerations in research as well as in teaching and learning. Thus, the significance of abductive reasoning requires being investigated not only within modeling but also within the inquiry methods observation, experimentation, and comparison (Figure 2).

In teaching and learning, hypothesis-driven empirical investigations with the help of different inquiry methods are often interpreted as deductive reasoning, whereas the students are also finding causes that explain a phenomenon and therefore are reasoning abductively. This frequently remains unrecognized in schools and in teacher education at university but can be seen as a resource for promoting creative thinking within scientific reasoning which should be strengthened by further empirical evidence.

Author Contributions: Conceptualization, A.U.z.B. and D.K.; methodology, D.K.; validation, D.K. and P.E.; analysis, P.E. and D.K.; investigation, A.U.z.B., P.E. and D.K.; resources, A.U.z.B. and D.K.; data curation, A.U.z.B., P.E. and D.K.; writing original draft preparation, A.U.z.B. and D.K.; writing review and editing, A.U.z.B., P.E. and D.K.; visualization, A.U.z.B., P.E. and D.K.; supervision, A.U.z.B. and D.K.; project administration, A.U.z.B. and D.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: The Probandes agreed to data use for research.

Data Availability Statement: The datasets are not publicly available.

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

References

1. Magnani, L. Model-Based and Manipulative Abduction in Science. *Found. Sci.* **2004**, *9*, 219–247. [CrossRef]
2. Nersessian, N.J. Model-Based Reasoning in Conceptual Change. In *Model-Based Reasoning in Scientific Discovery*; Magnani, L., Nersessian, N.J., Thagard, P., Eds.; Springer: Boston, MA, USA, 1999; pp. 5–22. ISBN 978-1-4615-4813-3.
3. Rocksén, M. The Many Roles of “Explanation” in Science Education: A Case Study. *Cult. Stud. Sci. Educ.* **2016**, *11*, 837–868. [CrossRef]
4. Giere, R.; Bickle, J.; Mauldin, R. *Understanding Scientific Reasoning*; Thomson: London, UK, 2006.
5. Rychen, D.S.; Salganik, L.H. *Key Competencies for a Successful Life and Well-Functioning Society*; Hogrefe Publishing: Göttingen, Germany, 2003; ISBN 978-1-61676-272-8.
6. Upmeier zu Belzen, A.; van Driel, J.; Krüger, D. Introducing a Framework for Modeling Competence. In *Towards a Competence-Based View on Models and Modeling in Science Education*; Upmeier zu Belzen, A., Krüger, D., van Driel, J., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 3–19. ISBN 978-3-030-30255-9.
7. Krüger, D.; Upmeier zu Belzen, A. Kompetenzmodell der Modellierkompetenz—Die Rolle abduktiven Schließens beim Modellieren. *Z. Für Didakt. Naturwiss.* **2021**. [CrossRef]

8. Kind, P.; Osborne, J. Styles of Scientific Reasoning: A Cultural Rationale for Science Education? *Sci. Educ.* **2017**, *101*, 8–31. [CrossRef]
9. Gouvea, J.; Passmore, C. ‘Models’ of versus ‘Models’ for: Toward an Agent-Based Conception of Modeling in the Science Classroom. *Sci. Educ.* **2017**, *26*, 49–63. [CrossRef]
10. Upmeyer zu Belzen, A.; Krüger, D. Modellkompetenz Im Biologieunterricht. *Z. Für Didakt. Naturwiss.* **2010**, *16*, 41–57.
11. Krüger, D.; Kauertz, A.; Upmeyer zu Belzen, A. Modelle und das Modellieren in den Naturwissenschaften. In *Theorien in der Naturwissenschaftsdidaktischen Forschung*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 141–157. ISBN 978-3-662-56319-9.
12. Clement, J. *Creative Model Construction in Scientists and Students: The Role of Imagery, Analogy, and Mental Simulation*; Springer: Dordrecht, The Netherlands, 2008; ISBN 978-1-4020-6711-2.
13. Johnson, T.R.; Krems, J.F. Use of Current Explanations in Multicausal Abductive Reasoning. *Cogn. Sci.* **2001**, *25*, 903–939. [CrossRef]
14. Peirce, C.S. *[Harvard] Lectures on Pragmatism*; Belknap, Harvard: Cambridge, MA, USA, 1978.
15. Krell, M.; Upmeyer Zu Belzen, A.; Krüger, D. Modellkompetenz im Biologieunterricht. In *Biologiedidaktische Forschung. Schwerpunkte und Forschungsgegenstände*; Sandmann, A., Schmiemann, P., Eds.; Logos: Berlin, Germany, 2016; pp. 83–102.
16. Krell, M.; Walzer, C.; Hergert, S.; Krüger, D. Development and Application of a Category System to Describe Pre-Service Science Teachers’ Activities in the Process of Scientific Modelling. *Res. Sci. Educ.* **2019**, *49*, 1319–1345. [CrossRef]
17. Gogolin, S.; Krüger, D. Students’ Understanding of the Nature and Purpose of Models. *J. Res. Sci. Teach.* **2018**, *55*, 1313–1338. [CrossRef]
18. Ke, L.; Schwarz, C. Using epistemic considerations in teaching: Fostering students’ meaningful engagement in scientific modeling. In *Towards a Competence-Based View on Models and Modeling in Science Education*; Upmeyer zu Belzen, A., Krüger, D., Driel, J., Eds.; Springer: Cham, Switzerland, 2019.
19. Kampourakis, K.; Niebert, K. Explanation in biology education. In *Teaching Biology in Schools: Global Research, Issues, and Trends*; Kampourakis, K., Reiss, M.J., Eds.; CRC Press: New York, NY, USA, 2018; pp. 237–248. ISBN 978-1-138-08798-9.
20. Park, J.H.; Lee, K.-H. How Can Mathematical Modeling Facilitate Mathematical Inquiries? Focusing on the Abductive Nature of Modeling. *Eurasia J. Math. Sci. Technol. Educ.* **2018**, *14*, em1587. [CrossRef]
21. Oh, P.S. How Can Teachers Help Students Formulate Scientific Hypotheses? Some Strategies Found in Abductive Inquiry Activities of Earth Science. *Int. J. Sci. Educ.* **2010**, *32*, 541–560. [CrossRef]
22. Oh, P.S. Features of Modeling-Based Abductive Reasoning as a Disciplinary Practice of Inquiry in Earth Science. *Sci. Educ.* **2019**, *28*, 731–757. [CrossRef]
23. Harman, G. The Inference to the Best Explanation. *Philos. Rev.* **1965**, *74*. [CrossRef]
24. Wirth, U. Die Phantasie Des Neuen Als Abduktion. *Dtsch. Vierteljahrsschr. Für Lit. Geistesgesch.* **2003**, *77*. [CrossRef]
25. Krems, J.; Johnson, T.; Kliegl, R. Kognitive Komplexität und abduktives Schließen. In *Strukturen und Prozesse Intelligenter Systeme*; Kluwe, R.H., Ed.; Deutscher Universitätsverlag: Wiesbaden, Germany, 1997.
26. Johnson, T.R.; Krems, J.; Amra, N.K. A computational model of human abductive skill and its acquisition. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*; Ram, A., Eiselt, K., Eds.; Erlbaum: Hillsdale, MI, USA, 1994.
27. Oh, P.S.; Oh, S.J. What Teachers of Science Need to Know about Models: An Overview. *Int. J. Sci. Educ.* **2011**, *33*, 1109–1130. [CrossRef]
28. Mahr, B. Information Science and the Logic of Models. *Softw. Syst. Model.* **2009**, *8*, 365–383. [CrossRef]
29. Agassi, J. Why there is no theory of models. In *Theories and Models in Scientific Processes. Proceedings of AFOS ‘94 Workshop, August 15–26, Madralin and IUHPS ‘94 Conference, August 27–29, Warszawa*; Herfel, W., Krajewski, W., Niiniluoto, I., Wójcicki, R., Eds.; Rodopi: Amsterdam, The Netherlands, 1995; pp. 17–26.
30. Gilbert, J.K.; Justi, R. *Modelling-Based Teaching in Science Education*; Models and Modeling in Science Education; Springer International Publishing: Cham, Switzerland, 2016; ISBN 978-3-319-29038-6.
31. Ritchey, T. Outline for a Morphology of Modelling Methods. *Acta Morphol. Gen. AMG* **2012**, *1*, 1012.
32. Mittelstraß, J. Anmerkungen zum Modellbegriff. In *Modelle des Denkens: Streitgespräch in der Wissenschaftlichen Sitzung der Versammlung der Berlin-Brandenburgischen Akademie der Wissenschaften am 12. Dezember*; Berlin-Brandenburgische Akademie der Wissenschaften: Berlin, Germany, 2005.
33. Stachowiak, H. *Allgemeine Modelltheorie*; Springer: Wien, Austria, 1973.
34. Mahr, B. Modelle Und Ihre Befragbarkeit Grundlagen Einer Allgemeinen Modelltheorie. *Erwäg. Wissen Ethik* **2015**, *26*, 329–342.
35. Harré, R. *The Principles of Scientific Thinking*; Macmillan: London, UK, 1970.
36. Passmore, C.; Gouvea, J.S.; Giere, R. Models in Science and in Learning Science: Focusing Scientific Practice on Sense-making. In *International Handbook of Research in History, Philosophy and Science Teaching*; Matthews, M.R., Ed.; Springer Netherlands: Dordrecht, The Netherlands, 2014; pp. 1171–1202. ISBN 978-94-007-7654-8.
37. Mahr, B. Ein Modell des Modellseins. In *Modelle*; Dirks, U., Knobloch, E., Eds.; Peter Lang: Frankfurt am Main, Germany, 2008.
38. Mäki, U. Models Are Experiments, Experiments Are Models. *J. Econ. Methodol.* **2005**, *12*, 303–315. [CrossRef]
39. Mahr, B. *On the Epistemology of Models*; De Gruyter: Boston, MA, USA, 2012; pp. 301–352. ISBN 978-3-11-025357-3.
40. Morrison, M.; Morgan, M.S. Introduction. In *Models as Mediators. Perspectives on Natural and Social Science*; Morgan, M.S., Morrison, M., Eds.; Cambridge University Press: Cambridge, UK, 1999.
41. Schurz, G. Patterns of Abduction. *Synthese* **2008**, *164*, 201–234. [CrossRef]

42. Popper, K. *Logik Der Forschung*; Mohr Siebeck: Tübingen, Germany, 2005.
43. Langlet, J. Kultur der Naturwissenschaften. In *Fachdidaktik Biologie*; Gropengießer, H., Harms, U., Kattmann, U., Eds.; Aulis: Hallbergmoos, Germany, 2016.
44. Greve, W.; Wentura, D. *Wissenschaftliche Beobachtung. Eine Einführung*; Beltz: Weinheim, Germany, 1997.
45. Upmeyer zu Belzen, A.; Krüger, D. Modelle Und Modellieren Im Biologieunterricht: Ein Fall Für Erkenntnisgewinnung. *Unterr. Chem.* **2019**, *171*, 38–41.
46. Bailer-Jones, D.M. Tracing the Development of Models in the Philosophy of Science. In *Model-Based Reasoning in Scientific Discovery*; Magnani, L., Nersessian, N.J., Thagard, P., Eds.; Springer: Boston, MA, USA, 1999; pp. 23–40. ISBN 978-1-4615-4813-3.
47. Knuuttila, T. Modelling and Representing: An Artefactual Approach to Model-Based Representation. *Stud. Hist. Philos. Sci. Part A* **2011**, *42*, 262–271. [CrossRef]
48. Grosslight, L.; Unger, C.; Jay, E.; Smith, C.L. Understanding Models and Their Use in Science: Conceptions of Middle and High School Students and Experts. *J. Res. Sci. Teach.* **1991**, *28*, 799–822. [CrossRef]
49. Justi, R.S.; Gilbert, J.K. Modelling, Teachers' Views on the Nature of Modelling, and Implications for the Education of Modellers. *Int. J. Sci. Educ.* **2002**, *24*, 369–387. [CrossRef]
50. Crawford, B.; Cullin, M. Dynamic Assessments of Preservice Teachers' Knowledge of Models and Modelling. In *Research and the Quality of Science Education*; Boersma, K., Goedhart, M., de Jong, O., Eijkelhof, H., Eds.; Springer: Dordrecht, The Netherlands, 2005; pp. 309–323. ISBN 978-1-4020-3673-6.
51. Priemer, B.; Eilerts, K.; Filler, A.; Pinkwart, N.; Rösken-Winter, B.; Tiemann, R.; Zu Belzen, A.U. A Framework to Foster Problem-Solving in STEM and Computing Education. *Res. Sci. Technol. Educ.* **2019**, *38*, 105–130. [CrossRef]
52. Krems, J.; Johnson, T. Integration of anomalous data in multicausal explanations. In *Proceedings of the 1995 Annual Conference of the Cognitive Science Society*; Moore, J.D., Lehman, J.F., Eds.; Erlbaum: Hillsdale, MI, USA, 1995.
53. Krüger, D.; Krell, M. Maschinelles Lernen mit Aussagen zur Modellkompetenz. *Z. Für Didakt. Naturwiss.* **2020**, *26*, 157–172. [CrossRef]
54. Clement, J.; Núñez-Oviedo, M.C. Abduction and Analogy in Scientific Model Construction. In Proceedings of the Annual Meeting of the National Association for Research in Science Teaching, Philadelphia, PA, USA, 23–26 March 2003.
55. Sturm, G. Abduktion. In *Methoden der Politikwissenschaft. Neuere Qualitative und Quantitative Analyseverfahren*; Behnke, J., Gschwend, T., Schindler, D., Schnapp, K.-U., Eds.; Nomos: Baden-Baden, Germany, 2006.
56. Göhner, M.; Krell, M. Preservice Science Teachers' Strategies in Scientific Reasoning: The Case of Modeling. *Res. Sci. Educ.* **2020**. [CrossRef]
57. Göhner, M.; Krell, M. Modellierungsprozesse von Lehramtsstudierenden Der Biologie. *Erkenn. Biol.* **2018**, *17*, 45–61.
58. Göhner, M.; Krell, M. Was ist schwierig am Modellieren? Identifikation und Beschreibung von Hindernissen in Modellierungsprozessen von Lehramtsstudierenden naturwissenschaftlicher Fächer. *Z. Für Didakt. Naturwiss.* **2021**. [CrossRef]
59. Reichertz, J. *Die Abduktion in der Qualitativen Sozialforschung: Über die Entdeckung des Neuen*; Springer: Wiesbaden, Germany, 2013; Qualitative Sozialforschung; 2., aktualisierte und erw. Aufl; ISBN 978-3-531-93163-0.
60. Lehrer, R.; Schauble, L. The development of scientific thinking. In *Handbook of Child Psychology and Developmental Science: Cognitive Processes*, 7th ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2015; Volume 2, pp. 671–714. ISBN 978-1-118-13678-2.
61. Heering, P.; Kremer, K. Nature of science. In *Theorien in der Naturwissenschaftsdidaktischen Forschung*; Krüger, D., Parchmann, I., Schecker, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2018.
62. Mayer, J. Erkenntnisgewinnung als wissenschaftliches Problemlösen. In *Theorien in der Biologiedidaktischen Forschung: Ein Handbuch für Lehramtsstudenten und Doktoranden*; Krüger, D., Vogt, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 177–186.
63. Gut-Glanzmann, C.; Mayer, J. Experimentelle Kompetenz. In *Theorien in der Naturwissenschaftsdidaktischen Forschung*; Krüger, D., Parchmann, I., Schecker, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2018.
64. Evans, J.S.B. Logic and Human Reasoning: An Assessment of the Deduction Paradigm. *Psychol. Bull.* **2002**, *128*, 978–996. [CrossRef]

Article

High School Students' Epistemic Cognition and Argumentation Practices during Small-Group Quality Talk Discussions in Science

Liwei Wei ^{1,*}, Carla M. Firetto ², Rebekah F. Duke ³, Jeffrey A. Greene ³ and P. Karen Murphy ¹

¹ Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University, University Park, PA 16802, USA; pkm15@psu.edu

² Mary Lou Fulton Teachers College, Arizona State University, Tempe, AZ 85287, USA; Carla.Firetto@asu.edu

³ School of Education, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; rduke23@live.unc.edu (R.F.D.); jagreene@email.unc.edu (J.A.G.)

* Correspondence: weiliwei@alumni.psu.edu

Abstract: For high school students to develop scientific understanding and reasoning, it is essential that they engage in epistemic cognition and scientific argumentation. In the current study, we used the AIR model (i.e., Aims and values, epistemic Ideals, and Reliable processes) to examine high school students' epistemic cognition and argumentation as evidenced in collaborative discourse in a science classroom. Specifically, we employed a qualitative case study approach to focus on four small-group discussions about scientific phenomena during the Quality Talk Science intervention (QT_S), where students regularly received explicit instruction on asking authentic questions and engaging in argumentation. In total, five categories of epistemic ideals and five categories of reliable processes were identified. Students demonstrated more instances of normative epistemic ideals and argumentative responses in the discussions after they received a revised scientific model for discussion and explicit instruction on argumentation. Concomitantly, there were fewer instances of students making decisions based on process of elimination to determine a correct scientific claim. With respect to the relationship of epistemic cognition to authentic questioning and argumentation, the use of epistemic ideals seemed to be associated with the initiation of authentic questions and students' argumentation appeared to involve the use of epistemic ideals.

Keywords: epistemic cognition; argumentation; science discussions; Quality Talk

Citation: Wei, L.; Firetto, C.M.; Duke, R.F.; Greene, J.A.; Murphy, P.K. High School Students' Epistemic Cognition and Argumentation Practices during Small-Group Quality Talk Discussions in Science. *Educ. Sci.* **2021**, *11*, 616. <https://doi.org/10.3390/educsci11100616>

Academic Editors: Moritz Krell, Andreas Vorholzer and Andreas Nehring

Received: 6 August 2021

Accepted: 30 September 2021

Published: 8 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

High school students must engage in the epistemic practices of science to develop their scientific understanding and reasoning [1]. It is not enough for them to read about and memorize the "facts" that have already been established by scientists. Science is an iterative, social process and the transmission of scientific facts from teacher to student does not do justice to the realities of science practices. Indeed, in the contemporary digital world where abundant unvetted information is easily created and spread [2], it is essential for students to develop reasoning skills and critically evaluate this information. Whether or not students choose to pursue a career in science, they must be armed with the ability to reason, problem-solve, as well as evaluate and justify arguments as they encounter scientific information in their daily lives. These abilities and practices are critical to navigating and effectively engaging in society. Rather than focusing on what science content students need to know, science education reforms have shifted the focus toward helping students understand how scientists observe the world and draw conclusions from their observations, leading to knowledge [3,4].

In line with this shift, over the past 50 years researchers have examined students' epistemic cognition, that is, how they acquire, understand, justify, change, create, and use knowledge [5–9]. The ways in which individuals conceptualize the fundamental nature of

how and what they know plays an important role in learning and acquiring knowledge [10], particularly for science education [9] and layperson scientific literacy [11]. For instance, in an attempt to further understand and model epistemic cognition in science, Reith and Nehring tested and confirmed the *ScieNo*-framework by examining the relationship between key scientific reasoning competencies and views on the nature of scientific inquiry (NOSI) [12]. Empirically, a recent meta-analysis has revealed that epistemic cognition interventions bolstered students' academic achievement in various ways and had the largest average effect size on argumentation among different types of academic achievement outcomes in the reviewed studies ($ES = 1.047, p < 0.001$) [13].

Indeed, both epistemic cognition and argumentation are core to scientific reasoning. According to Osborne, scientific reasoning abilities require "a meta-level knowledge of science and the epistemic features of science," (p. 274) which are necessary for learners to understand why certain scientific claims are warranted [14]. Indeed, Chinn and Sandoval argued that scientific reasoning requires that students acquire, understand, and use scientific practices and norms, which include various facets of epistemic cognition such as reliable processes for knowing and the skills to generate arguments that support knowledge claims [15]. In addition, the emphasis on fostering scientific reasoning skills requires that students engage in argumentation during which they construct and evaluate scientific models through evaluation [4]. During this process, students are expected to provide justifications based on evidence to support or refute claims, which is central in scientific reasoning [16]. In line with Upmeier zu Belzen and colleagues in this Special Issue, "modeling is a prominent style of scientific reasoning" [17] (p. 495).

Delving into these key facets of scientific reasoning, researchers have also identified a close relationship between epistemic cognition and argumentation. Indeed, epistemic cognition influences and supports argumentative reasoning; more complex or developed epistemic beliefs are related to better argumentative reasoning skills (e.g., production and evaluation of arguments; [18–22]). In turn, argumentation is often emphasized as a part of epistemic cognition interventions [23,24] and may also serve to promote epistemic cognition. As a case in point, Jordanou and Constantinou found that 11th-grade students who participated in evidence-focused, argumentative discourse activities in a Web-based learning environment increased their use of scientific evidence in their electronic dialogs with a peer, used more evidence to weaken the opponent's claims, and made explicit references to the source of evidence, whereas the comparison counterparts did not exhibit such improvements [25]. This suggests that students developed a more advanced epistemological understanding in science after engaging in sustained argumentation. However, given the complexity of epistemic cognition, scientific argumentation, and their interaction, more research is needed to more clearly delineate the relations between collaborative argumentation and students' epistemic cognition.

Informed by this body of literature, we investigated the epistemic cognition practices of high school students during small-group science discussions over the course of a year-long intervention designed to develop students' argumentation and discourse skills. We chose to leverage small-group discussions to examine students' epistemic practices in science, given the established literature documenting positive effects of small-group discussions on students' scientific argumentation and critical-analytic thinking [26–28] as well as the aforementioned evidence regarding the relationship between argumentation and epistemic cognition in science [25]. Specifically, we conducted a qualitative case study of a small group of students engaging in four science discussions, with the goal of better understanding how epistemic cognition and scientific argumentation manifest and interact with one another, particularly as students learn more about argumentation and discourse. This research contributes understandings of how best to incorporate small-group discourse into science classrooms, engage students in epistemic practices of science, and prepare students to think critically and analytically about the scientific information they encounter in their daily lives.

1.1. Theoretical and Explanatory Framework: AIR Model of Epistemic Cognition

To examine and analyze students' epistemic cognition as reflected in their small-group discourse in science classrooms, we employed Chinn and colleagues' AIR model of epistemic cognition (i.e., Aims and values, epistemic Ideals, and Reliable processes) [29,30] as the theoretical and explanatory framework of the current study. Chinn and colleagues proposed that epistemic cognition is comprised of three components [29,30]. The first component, *epistemic aims*, consists of goals related to the pursuit of epistemic ends or products, such as knowledge, understanding, explanations, true beliefs, scientific models, or rational arguments. For example, students who adopt the aim of summarizing explanations deemed normative by the field (i.e., knowledge) would necessarily engage in a given task differently from those whose epistemic aim was to achieve deep understanding of those explanations, such as the scientific reasoning that underlies such explanations [31,32].

The second component, *epistemic ideals*, represents the criteria or standards used to evaluate epistemic products, which are discipline-specific, context-specific, and even topic-specific. Students use epistemic ideals as justification for the adequacy of the epistemic products they construct or to evaluate the epistemic products of other individuals. For example, a scientific claim's adequacy as an epistemic product can be judged by how well it adheres to various science-based epistemic ideals, such as its fit with prior knowledge. Chinn and colleagues proposed five broad categories of epistemic ideals: (a) specification of the internal structure of an epistemic product, (b) connection to and coherence with other knowledge, (c) present and future connections to empirical evidence, (d) credibility of testimony, and (e) coherency and how well it has been communicated [30]. For example, when considering models of a scientific phenomenon, students may hold the epistemic ideal that "good models fit all the evidence," or "good models are parsimonious." Likewise, when evaluating a scientific argument, students may hold the epistemic ideal that "strong evidence addresses core parts of the model," or "good arguments are clearly communicated."

Reliable epistemic processes, the third component of the AIR model, are the methods by which knowledge and other epistemic products are constructed [29,30]. Reliable processes, strategies, and practices are those that consistently result in epistemic products that meet epistemic aims. Classification as to whether a process is reliable and appropriate depends largely on the discipline and context; they are often contingent upon the circumstances in which they are enacted, although certain processes are near-universally viewed as less reliable than others (e.g., relying on hearsay). In science, controlled experimentation and rigorous observation are often endorsed as reliable processes, under particular conditions. Observation may be a reliable process when a person uses it to visually count a small number of people in a room, but it becomes much less reliable when counting people in a crowd of thousands. As argued by Chinn et al., a critical part of epistemic cognition relates to the people's schemas about the conditions under which processes can be considered reliable, and these schemas can be used in different ways [30]. According to Chinn et al., individuals may use the schemas to guide their actions, that is, to *enact* a reliable process [30]. For example, a student may conduct a well-designed scientific experiment to collect data as evidence to support a claim. Second, individuals may use the schemas to *evaluate* the processes used by others, such as by judging whether a specific method is viable to generate an accurate understanding of a scientific phenomenon. Third, individuals may use the schemas to express *metacognitive* beliefs about how to produce reliable epistemic products. For instance, an individual may explain what needs to be considered when evaluating a scientific argument.

Chinn and colleagues' model has been used as a framework to examine learners' epistemic cognition while engaged in scientific inquiry [24,33]. For instance, Herrenkohl and Cornelius examined the argumentation practices of fourth- and fifth-grade students and teachers to assess the epistemic thinking that emerged during instructional activities such as whole-class discussions and small-group discussions [24]. The researchers coded whole-class discourse for argumentation and categorized the emergent code clusters into

the components of Chinn et al.'s model of epistemic cognition. In the present study, we also employed the AIR model as the theoretical and explanatory framework. Specifically, we conceptualized epistemic cognition based on the AIR model. Further, we used the AIR model to identify and analyze student discourse to deepen understanding of how epistemic ideals and reliable processes occur and interact together with argumentation in student discourse about science in small-group discussions. The goal was to gather a better sense of the criteria students use to form scientific arguments and to better understand the relationship between students' practices of argumentation and the epistemic criteria they hold and apply in science classrooms.

1.2. Interplay of Argumentation and Epistemic Cognition in Science

As addressed above, the AIR model aligns well with contemporary research and theory on argumentation. Argumentation—the process through which knowledge claims are asserted and justified through supporting reasons and evidence—is part of the foundation for the development and progression of scientific knowledge [34–37]. Thus, when conducted in ways that adhere with scientific normative practices, argumentation can be considered a reliable epistemic practice in science. Scientists advance knowledge in their field by endorsing normative epistemic ideals such as that a scientific argument needs to be supported by evidence and connected to prior theories (i.e., coherence with evidence or normative disciplinary knowledge). Subsequently, scientists must establish a convincing argument and communicate it to the broader scientific community. Their argument is subject to critical evaluation by their peers, who can question aspects of the argument and make counterarguments. The goal of reasoned argumentation is thus to come to a rational conclusion about which claims to accept or which actions to take [38].

As students engage in argumentation as an epistemic process, they are also likely to develop their epistemic understanding of science [25]. When students engage in collaborative argumentation, their arguments are also open to evaluation by others, who can examine the provided justification and accept or reject the purported claims [39]. During this process, alternative positions can be considered as well. Specifically, an individual can engage in written argumentation independently by articulating their own viewpoints and providing reasoning and evidence in support of their claim as well as considering multiple perspectives and counterarguments to their position. However, students can also engage in oral argumentation collaboratively and dialogically. During oral argumentation, students benefit from listening to others, processing and evaluating others' arguments, similar to what scientists do in their own practice. As a result, engaging students in argumentation helps them to understand the processes behind science and to develop a deep understanding of how knowledge develops in the scientific discipline [40], which subsequently advances their science learning [3,41–43].

Epistemic ideals, on the other hand, guide the kinds of reasons and evidence used in the scientific arguments constructed by scientists and students. In science, there are disciplinary standards (i.e., epistemic ideals) regarding the ways in which argumentation (e.g., evidence or connection to other theories) is used in knowledge building [3]. These disciplinary standards are the accepted guidelines by which the community justifies and evaluates knowledge, as well as the processes used to produce knowledge [3]. As a result, argumentation involves a deliberation on the epistemic status of knowledge claims [44]. For instance, in science, claims that adhere to scientific evaluative criteria (e.g., supported by evidence or fit with prior theories) are given predominant epistemic status over claims that do not meet these criteria. In the coordination of claims, reasons, and evidence, one's epistemic cognition becomes pivotal. Absolutist, multiplist, and naïve views of knowledge and knowing provide few guides as to what should and should not be considered a valid knowledge claim [45], whereas when students adopt an evaluativist perspective and more normative beliefs, they are more likely to utilize disciplinary norms to evaluate arguments and consider whether to accept or refute the arguments.

Therefore, the epistemic ideals students hold will guide the kinds of reasoning, evidence, and arguments they bring forward and the type of disciplinary standards they use to evaluate the presented arguments. Empirical evidence shows that students' epistemic cognition influences how they evaluate and construct scientific arguments [46]. Students with more advanced epistemological understanding engage in more critical evaluation [47]. They are better able to identify informal reasoning fallacies in flawed arguments [48] and produce higher-quality written arguments of their own [18]. Nussbaum and colleagues examined the transcripts of paired students' online argumentation discussions and found that students with less advanced epistemological understanding were less critical of their partner's arguments [19]. Also, these students did not acknowledge inconsistencies within arguments, when compared to students who held more advanced views. The more advanced students provided counterarguments, brought forth more content into their argumentation, and noted the need for more information. Students with more advanced epistemic perspectives were also more willing to engage in argumentation than peers with more naïve perspectives [19,49]. Notably, there is literature suggesting this relationship could be bi-directional. When students engage in dialogic argumentation and demonstrate their knowledge of the argumentation norms in science, they reveal an improvement in their epistemic understanding [50]. Given the strong alignment between models of argumentation and epistemic cognition (e.g., AIR model), there is a need for research on how to construct argumentation instruction in ways that help students refine their epistemic understanding of science, which necessarily includes normative scientific aims, ideals, and reliable processes.

1.3. Using Quality Talk Science (QT_S) as a Potential Approach to Enhance Argumentation and Examine Epistemic Cognition

As stressed in prior research, the kind of classroom intervention found to be effective for promoting epistemic cognition often involves teachers' creating and supporting an open space where small groups of students can co-construct and challenge arguments about domain-specific problems [2,51]. Further, within the context of an open participation space, the type of task assigned to students may also influence their performance. For example, a well-defined, open-ended, and challenging task provides more opportunities for students to utilize multiple strategies and can help promote generalization, argumentation, and higher-order thinking [52,53]. The variable nature of open-ended tasks also stimulates conversations among students to allow for a negotiation of meaning and understanding of the domain knowledge [54].

In this study, we examined students' oral discourse in science during an intervention called Quality Talk Science (QT_S), a teacher-facilitated, small-group, discourse intensive approach that aims to promote students' critical-analytic thinking and high-level comprehension about scientific models and phenomena [27,55]. Similar to the aforementioned characteristics of successful interventions that promote epistemic cognition, during QT_S teachers receive a series of professional development workshops to become familiar with the pedagogical principles. Specifically, these pedagogical principles outline the need for students to take on the interpretative authority of the discussion. To achieve this, teachers gradually release control of the discussion to students, such that students increase their responsibility participating in productive discourse about scientific content, searching for the underlying arguments and assumptions (i.e., epistemic engagement) [56]. Further, as part of these pedagogical principles, teachers provide explicit instruction to students with guided practice on how to generate thought-provoking, open-ended questions (i.e., authentic questions, AQ) and respond to those questions using argumentation. These student-initiated authentic questions and argumentation responses serve as indicators of high-level comprehension, as students critically and reflectively engage with scientific text or content. As an essential part of QT_S, students engage in regular small-group discussions where they are expected to evaluate scientific models related to various scientific phenomena. Teachers facilitate these discussions using appropriate teacher discourse moves such as marking or modeling discourse elements indicative of productive talk [57].

QT_S has both theoretical and empirical underpinnings as a branch of the broader Quality Talk (QT) framework [58]. QT was derived from a systematic review of text-based discussion interventions in language arts [26] and was adapted for use in high-school science classrooms [27]. The most effective parts of multiple approaches to discussion were combined into one approach designed to bolster students' high-level comprehension and critical-analytic thinking of text. It is rooted in rich theoretical underpinnings including cognitive, sociocognitive, sociocultural, and dialogic perspectives on teaching and learning [59].

Accumulating empirical research on the QT approach has evidenced positive impacts on improving students' discourse and argumentation in science, literacy, English language learning, and mathematics [27,60–62], as well as in different cultural contexts [63–65]. As a case in point, we conducted a quasi-experiment in which high school chemistry and physics teachers implemented QT_S in their classrooms over a school year [27]. The critical-analytic thinking and argumentation in the discourse of students engaging in QT_S improved dramatically from pre-test to post-test. Over time, QT_S students asked more questions that provoked deeper levels of cognitive processing and responses [27]. In contrast, students in the comparison classroom did not evidence these changes to the same degree. At the end of the school year, QT_S students produced many more well-supported responses with reasoning and evidence, and challenged and built on others' arguments more frequently. Such indicators were not present in the pre-test discussions, nor were they present in the post-test discussions of the students in the comparison classroom. Comparable results have been shown across varying grades, content areas, and contexts. For example, in language arts classrooms, fourth-grade students who participated in QT discussions evidenced increases in students' basic- and high-level comprehension [60] as well as students' written argumentation after receiving writing instruction as part of the QT intervention [66].

In sum, QT_S has shown promise as a way to foster scientific practices that involve argumentation and understanding via small-group discussion, and it aligns well with instruction on epistemic cognition in science. However, less is known about the epistemic ideals that students use in scientific discourse as they generate arguments and how they consider reliable processes while understanding scientific phenomena.

1.4. The Present Study

In this qualitative case study, we examined how high school students engaged in small-group discussions about scientific models and phenomena with a particular focus on how students' epistemic cognition and argumentation were evidenced across a set of discussions. We used the AIR model as the theoretical and explanatory framework from which we identified and analyzed the epistemic ideals and reliable processes students used while constructing arguments and evaluating scientific models. This study contributes to the extant literature in three ways: (a) our methodological approach allowed us to gather evidence of epistemic cognition and argumentation as enacted in students' oral discourse rather than via self-reports, (b) the AIR model enabled us to capture the criteria students used to evaluate scientific arguments while also contributing to the emerging body of literature using the framework to analyze collaborative argumentation discourse [24,33], and (c) the use of the QT_S discussion approach contributed to examining the relationship between epistemic cognition and argumentation as well as informing instructional implications for promoting scientific argumentation and epistemic cognition in science classrooms. Our research questions were:

- RQ1.** What types of epistemic ideals, reliable processes, and argumentation do students invoke while engaging in small-group, QT_S discussions?
- RQ2.** How do students' epistemic cognition and argumentation vary based on contextual factors of the discussions (i.e., model format and explicit instruction)?
- RQ3.** How does students' epistemic cognition relate to authentic questioning and argumentation during QT_S discussions?

2. Methods

2.1. Participants and Study Design

Within the context of a larger National Science Foundation grant, four teachers from one public high school in the northeastern United States implemented QT_S in their 10th-through 12th-grade chemistry and physics classes over an entire academic year. Students in the school were predominantly Caucasian (i.e., 91%), and over half of the students were from economically disadvantaged families (i.e., 49% qualified for the Free Lunch Program and 8% qualified for the Reduced-Price Lunch Program). The school was situated within a small city in a rural setting. The student population was highly transient; almost half of the participants who enrolled in the study at the start of the school year changed school districts over winter break.

For this qualitative case study, one group of all female students ($n = 6$) from the AP chemistry class was selected for analysis. Although students in the class were split into four discussion groups, we elected to examine the discourse from one of the small groups so that we could conduct the depth of qualitative analysis necessary to explore our research questions. We identified the best fitting group for analysis based on two primary selection criteria: (a) a group where the teacher was not present for the QT_S science lesson discussions and (b) a group with students who had high rates of attendance and a full year of participation. These selection criteria allowed us to identify the group that would give us the best sense of students' epistemic cognition and scientific argumentation without the influence of the teacher or the variability in group composition (e.g., shifts in group dynamics due to student absences). Finally, it is important to note that in this qualitative study, we emphasized ecological validity over external validity. That is, the study examined student learning in an authentic science classroom. Therefore, our research design does not warrant causal claims or generalizations from our findings.

2.2. QT_S Intervention

The key components of the QT_S intervention included the delivery of QT_S discourse lessons and QT_S catalyst, QT_S science lessons, QT scientific model handouts for QT_S discussions, and QT_S discussions across one academic year (see Table 1 for schedule and timeline), which are introduced in the following sections, respectively.

Table 1. Timeline of Monthly Cycles with QT_S Discourse Lessons and QT_S Science Lesson Topics.

Month	QT _S Discourse Lesson Content	QT _S Science Lesson Topic
Emphasis on Asking Open-Ended Questions (Fall)		
September	Authentic Questions	Airbags *
October	Question Types	Soap Bubble
November	Question Types	Nuclear Fission *
Emphasis on Argumentation (Spring)		
January	Components of an Argument	Thin Films *
February	Evaluating Evidence and Reasoning	Hot Packs *
March	Counter-Argument	Tesla Coil

Note. * Denotes discussion analyzed as part of this study's data.

2.2.1. QT_S Discourse Lessons and QT_S Catalyst

The six QT_S discourse lessons were shared with the teacher during the initial and ongoing professional development workshops. For each discourse lesson, the teacher was provided with a set of slides to present in class as well as a corresponding lesson plan. The first three discourse lessons focused on different types of authentic questions that students could generate and ask in their discussions (e.g., speculation questions, connection questions, or high-level thinking questions) and were delivered in fall. The last three discourse lessons were delivered in spring and were focused on teaching students about argumentation components (i.e., claim, reasoning, and evidence), the evaluation of

evidence and reasoning, as well as challenge, alternative argument, and counterargument. Students were not only introduced to the definition of each necessary argumentation component, but they were also provided with guidelines (i.e., relevance, credibility, and accuracy) on evaluating evidence and quality of reasoning. All discourse lessons included descriptions of concepts as well as realistic examples of these concepts illustrated through discussion transcripts and/or videos [58].

Students were also provided with a QT₅ catalyst worksheet to correspond with the discourse lessons and prepare students for the QT₅ discussions. In the fall semester, the QT₅ catalyst focused on different types of authentic questions in alignment with the QT₅ discourse lessons (Figure 1a). The fall QT₅ catalyst provided space for students to record their authentic questions about the model, readings, and demonstration in preparation for discussion. In spring, students were provided with a QT₅ catalyst that centered on argumentation in alignment with the discourse lessons focused on argumentation. In addition to providing space for recording authentic questions, the spring QT₅ catalyst used visual representations of each argumentation component to facilitate the discussion and help students think about the model for discussion regarding the evidence and reasoning for each claim (Figure 1b).

(a)

(b)

Figure 1. Examples of QT₅ Catalysts. (a) The top QT₅ catalyst was used in fall with a focus on recording students’ authentic questions; (b) The bottom QT₅ catalyst was used in spring with a focus on argumentation components.

2.2.2. QT₅ Science Lessons

Paired with each of the six discourse lessons, the teacher also taught a QT₅ science lesson over three consecutive class periods. QT₅ science lessons were co-created with teachers and content area experts to provide rich opportunities for students to engage in discussions around disciplinary core ideas in science in alignment with the Next Generation Science Standards (NGSS). Each lesson was centered around an essential question related to a scientific phenomenon (see Tables 1 and 2 for details).

Table 2. QT₅ Science Lesson Details.

QT ₅ Science Lesson Topic	Essential Question	Science Concepts	Class Demonstrations
Airbags	How does the inflation and deflation of the airbag prevent injury?	Newton's Laws of Motion, Kinetic Theory of Gases, Acceleration, Velocity, Force, Diffusion	<ul style="list-style-type: none"> • Video demonstration of a crash test with and without an airbag • Video demonstration of an airbag deployment and deflation in slow motion
Nuclear Fission	How does nuclear fission create explosions?	Fission, Strong force, Nucleons, Nuclides, Neutrons, Protons, Electrons, Binding Energy, Electrostatic Forces, Radiation, Isotopes, Stability	<ul style="list-style-type: none"> • Video demonstration of chain reactions • Video demonstration of an explosion from 100 tons of TNT • Video discussion of the Manhattan Project Trinity Test
Thin Films	What causes the appearance of multiple colors in a layer of colorless nail polish when it is observed under white light?	Destructive Interference, Constructive Patterns, Refraction, Young's Experiment, Absorption, Scatter, Diffraction, Reflection, Light Dispersion, Miscible, Immiscible, Density	<ul style="list-style-type: none"> • Hands-on, thin film rainbow paper experiment
Hot Packs	Why does clicking the disk in a reusable hot pack result in the release of heat?	Phase Change, Exothermic, Endothermic, Entropy, Energy, Activation Energy, Potential and Kinetic Energy, Enthalpy	<ul style="list-style-type: none"> • Video demonstration of a hot pack activating in slow motion • Hands-on, reusable hot pack for each group

On the first day of the QT₅ science lesson, students were introduced to the essential question and observed demonstrations of the phenomenon (i.e., hands-on activity or video). After the demonstrations, the teacher introduced a handout that contained multiple models/claims to explain the scientific phenomenon that students observed in the demonstration (Figure 2). During and after the demonstrations, students generated and wrote down authentic questions about the phenomenon or their thinking about each claim in the scientific model on their QT₅ catalyst worksheet (Figure 1). Taken together, students were provided with multiple exposures to the scientific phenomenon and related material (e.g., demonstrations or scientific readings) in order to promote the likelihood that students possessed the necessary foundational understanding related to the phenomenon prior to the discussion. This approach also allowed multiple opportunities for students to develop a variety of rich authentic questions and engage with the scientific model. On the second day, students brought their QT₅ catalyst, readings, and the provided scientific model/claims to their QT₅ discussions to talk about the scientific model and related scientific phenomenon. On the third day of the science lesson after students conducted a QT₅ discussion, the teacher reviewed the student evaluations of the presented models or claims via a whole-class discussion toward the normative model and addressed any

misconceptions in student responses to help them understand the normative scientific model in class.

Quality Talk Name: Teacher's Name:
Date: Period:

Scientific Models ~ Airbags

Directions: Read each scientific model. Pay close attention to how the models describe the materials and structures involved, the scientific process(es), observations, and the atomic and molecular levels of activities. If you have a question about any of the models, record it on your Quality Talk Catalyst worksheet.

Model 1
The airbag inflates through a rapid chemical reaction, generating nitrogen gas (N₂). This gas fills a nylon or polyamide bag at a velocity of 150 to 250 miles per hour. At the same time that the airbag is inflating, the nitrogen gas is escaping from very small holes. The airbag inflates in order to provide a force in the equal and opposite direction of the driver's motion. Equal forces in opposite directions cancel each other out, eliminating any forward force that might cause injury to the driver. After the initial inflation of the airbag due to a collision, the airbag deflates because air does not take up space. The deflation of the airbag increases the space in front of the driver for easy removal of the person.

Model 2
The airbag inflates through a rapid chemical reaction, generating nitrogen gas (N₂). This gas fills a nylon or polyamide bag at a velocity of 150 to 250 miles per hour. At the same time that the airbag is inflating, the nitrogen gas is escaping from very small holes. As the airbag deflates, it slows the forward velocity of the driver. As the nitrogen gas escapes through the small holes, it increases the time it takes for the driver to hit the steering wheel. The slower forward velocity prevents the driver from any injuries. After the initial inflation of the airbag due to a collision, the airbag deflates because air does not take up space. The deflation of the airbag increases the space in front of the driver for easy removal of the person.

Model 3
The airbag inflates through a rapid chemical reaction, generating nitrogen gas (N₂). This gas fills a nylon or polyamide bag at a velocity of 150 to 250 miles per hour. At the same time that the airbag is inflating, the nitrogen gas is escaping from very small holes. As the airbag deflates, it slows the forward velocity of the driver. As the nitrogen gas escapes through the small holes, it increases the time it takes for the driver to hit the steering wheel. The slower forward velocity prevents the driver from any injuries. The diffusion of the nitrogen gas aids in slowing down the movement of the driver.

Model 4
The airbag inflates through a rapid chemical reaction, generating nitrogen gas (N₂). This gas fills a nylon or polyamide bag at a velocity of 150 to 250 miles per hour. At the same time that the airbag is inflating, the nitrogen gas is escaping from very small holes. The airbag inflates in order to provide a force in the equal and opposite direction of the driver's motion. Equal forces in opposite directions cancel each other out, eliminating any forward force that might cause injury to the driver. The diffusion of the nitrogen gas aids in slowing down the movement of the driver.

(a)

Thin Film Model by: Alice Class: Mrs. Brady pd4.

Claim 1 When a drop of nail polish is dropped onto a water surface the lower density of the nail polish and the molecular attraction of the molecules prevent it from mixing with the water.

Claim 2 The white light refracts off the film, splitting into waves of different colors.

Claim 3 The different color patterns in the film that we see are due to the light waves bouncing off each other as they travel to our eyes.

(b)

Figure 2. Examples of handouts for scientific models/claims. (a) The one on the left was used in fall with four models for students to choose from; (b) The one on the right was used in spring and resembles a student-generated model with three different claims.

2.2.3. Scientific Model Handouts for QT_S Discussions

The scientific model handouts were an integral part of the QT_S science lesson, as they provided a framing for alternative scientific models related to the phenomena and also served to guide the discussions. In line with Schwarz et al. [67,68], scientific models are considered “tools for predicting and explaining” scientific phenomena, and scientific models can “change as understanding improves” [67] (p. 632). Having students engage in modeling practices is conducive to developing their epistemic understanding as well as their capacity for constructing and evaluating knowledge in science [69]. Therefore, in the current study, students were afforded the opportunity to evaluate and revise these models during QT_S discussions as part of their learning about various scientific phenomena.

In fall, the handout consisted of four different models of the given phenomena. Each model had a collection of claims and there were overlapping claims across the four models. Among the four models, one of them contained all correct claims and the remaining three had one or more incorrect claims (Figure 2a). However, the teacher and students reported that these models were too simplistic. Once students identified one model that they believed to be correct, they no longer considered the remaining models. As a result, we revised the handout. In spring, a single model was presented, which included three claims that jointly explained the phenomena (Figure 2b). The models and claims were hand-drawn and formatted to appear as if they were student-generated work rather than an authoritative source (e.g., a textbook figure). The three claims addressed different aspects of the model and did not have any overlapping components. Students were told that any

number of the claims were potentially correct or incorrect, and their task was to provide reasoning and evidence regarding the veracity of each claim. For incorrect claims, students were asked to generate a correct claim with appropriate reasoning and evidence.

2.2.4. QT_S Discussions

The small-group discussions took place on the second day of the QT_S science lesson. Given logistical and time constraints, it was not possible for the teacher to facilitate four discussions in one day while still allowing each group enough time to engage sufficiently in discussions (i.e., at least 15 min). Thus, the teacher organized the class so that two groups discussed for the first half of the class, while the other two groups worked independently on classwork, and then they switched for the second half of class. The teacher facilitated one small-group discussion in each half, while the other group engaged in a discussion without a facilitator. Discussions lasted approximately fifteen minutes and naturally unfolded in two portions: (a) discussing the answer to the essential question presented in the lesson, also called the model-based portion of the discussion and (b) engaging in the discussion about related scientific content guided by student-initiated questions, also called the open-ended portion of the discussion.

In response to the essential question about the provided model, students discussed the different models or claims with respect to which were correct, which were incorrect, and why. Notably, there was no single answer to these questions and there existed multiple ways to address these questions by referring to various pieces of evidence from the provided readings or demonstrations. During the model-based portion of the QT_S discussions, students focused on one specific epistemic aim: determining whether the model or claim was scientifically sound. To achieve this epistemic aim, students needed to evaluate and analyze the scientific credibility of the provided models or claims using reasoning and evidence.

After students concluded their discussion around the essential question and reached a conclusion regarding the scientific model, they began the open-ended portion of the discussion. When students engaged in the open-ended portion of the discussion, a singular, central epistemic aim was not evident. This open-ended portion revolved around asking and answering student-generated authentic questions. These two distinct parts emerged from the flow of the discussion across all discussions, but occasionally, students would briefly return to reconsidering the essential question as it related to a student-generated authentic question they were discussing. Importantly, this split between the scientific model-based portion and the open-ended portion of the discussion was not invoked by the teacher nor controlled by the researchers.

2.3. Procedures

Along with a cohort of teachers participating in the larger grant-funded study, the chemistry teacher participated in a series of initial and ongoing professional development workshops, where they learned about the QT_S approach and how to implement it in their classroom with researcher-provided materials (e.g., QT_S discourse lessons, QT_S science lessons, or materials for hands-on activity). They also received regular one-on-one coaching sessions with QT_S coaches to support high-fidelity QT_S implementation (e.g., successful delivery of QT_S discourse lessons, QT_S science lessons, or implementation of QT_S discussions).

Each month the teacher engaged in a cycle (see Table 1 and Figure 3) whereby they: (a) presented a QT_S discourse lesson to teach aspects of authentic questions and argumentation, (b) implemented a QT_S science lesson about a disciplinary core idea, (c) conducted small-group discussions based on the disciplinary core idea science lesson with two groups being teacher-facilitated and two groups being student-led, which were both video- and audio-recorded, and (d) reviewed student evaluation of the scientific model(s) through a whole class discussion and presented the normative scientific model. In addition, the teacher also (e) conducted a second set of small-group discussions based on a chemistry

lesson of their choice. In this teacher-choice science discussion, the teacher used a structure similar to the QT_S science lessons but without a scientific model. This gave students an opportunity to engage in additional QT_S discussions while also allowing the teacher to facilitate the groups that were previously student-led. Thus, by alternating the discussions, the teacher was able to facilitate all four groups within each cycle. Over the academic year, this monthly cycle repeated six times. In this study, we examined four of the QT_S discussions conducted by one student-led group.

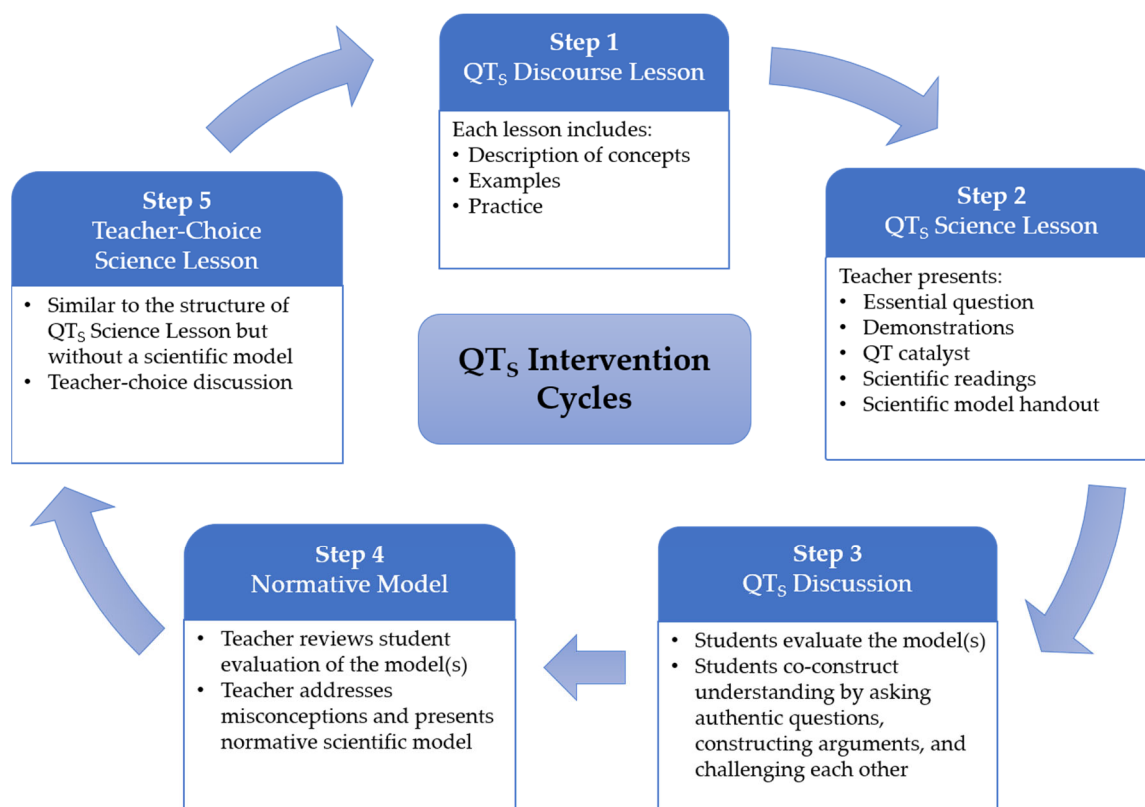


Figure 3. A Flowchart of the QT_S Intervention Procedures (One Cycle).

2.4. Qualitative Coding

2.4.1. Epistemic Cognition Coding

The coding for epistemic cognition was conducted through iterative cycles to ensure consistency and coherence across the two portions of each QT_S discussion (i.e., model-based portion and open-ended portion). Prior literature regarding epistemic criteria for models and arguments [70] and Chinn and colleagues' AIR model [30] framed our initial, open coding of students' discourse. After separately coding the discourse, the authors met to discuss the emergent codes in the data and our reasoning for each code. Over multiple coding cycles, we refined our coding scheme and resolved all discrepancies through discussion. Given the two portions of each QT_S discussion had different epistemic aims, coding was conducted first for the model-based portion of the discussion and then replicated for the open-ended portion in order to maintain greater consistency and to avoid drift in coding for each portion of the QT_S discussion. Throughout the coding process, the researchers wrote analytic memos [71] to document reflections and insights regarding the coding process. It is worth noting that the authors did not code for epistemic aims in the coding of epistemic cognition. This was because there was one clear epistemic aim for the model-based portion of the discussion, which was to determine a correct model, and no clear, central epistemic aim for the open-ended portion of the discussion during which students answered their authentic questions related to the scientific phenomenon

in general. Through this iterative, multi-cycle process, the authors developed two sets of epistemic cognition codes (i.e., epistemic ideals and reliable processes) with five categories of epistemic ideals and five categories of reliable processes. While coding for the reliable epistemic processes used in the discussions, the researchers also noted the ways of how students used the schemas that specify these processes, such as whether each reliable epistemic process was enacted, evaluative, or metacognitive [30]. See Tables 3 and 4 for code descriptions and examples.

Table 3. Descriptions of Epistemic Ideals Codes with Examples.

Epistemic Ideal Code	Code Description ¹	Example ²
Coherence with normative disciplinary knowledge	The explanation is consistent with known scientific knowledge from an authoritative source. This includes information from the provided texts or any scientific knowledge they apply (e.g., prior knowledge from another science class).	<p style="text-align: center;">Nuclear Fission</p> <p>Emma: I had a question about the decaying. Like, why do they decay? Is it to become more stable? Aria: Yeah, because remember how it [the article] says, like, the stability of an isotope is dependent on the ratio of the number of neutrons and protons. Well, it was stable before, but if you added another neutron it's unbalanced. So, it wants to be stable, when it's not, it's just going to decay. {EI: Coherence with NDK}</p>
Coherence with personal experience	The explanation is consistent with a situation that they have personally experienced.	<p style="text-align: center;">Hot Packs</p> <p>Grace: 'Cause, like, in the [article] it said 130 degrees Fahrenheit, but do you think, some . . . like, I doubt that [the reusable hot pack] got that hot . . . {EI: Coherence with PE} Aria: Yeah, it's not that hot. Fifty-four, um . . . In the summer, the temperature is roughly, like, 38 degrees Celsius, so really, that's not that hot. {EI: Coherence with PE}</p>
Coherence with the personal experience of a layperson	The explanation is consistent with personal experiences of others that they were told or heard about (e.g., friend, family, coworker).	<p style="text-align: center;">Airbags</p> <p>Aria: Well, say, like I told you earlier, my dad was in an accident, like, a year ago. A car hit him from behind, so his face physically went forward, and the airbag popped out, so he went backwards again. So his neck was sprained and he couldn't move his neck around for a month. So is that injury really necessary? {EI: Coherence with PE of a layperson}</p>
Coherence with prior knowledge (other)	The explanation is consistent with other prior knowledge that is not from their own personal experience and is not scientific knowledge they know.	<p style="text-align: center;">Airbags</p> <p>Chloe: I think in the demonstration they were, like, wearing, like, a lap belt and, like, when they, like, came up, like, the airbag hit them and they went back down. But, like, without the seatbelt they would probably go over the airbag, you know what I mean and then, like, crash. {EI: Coherence with PK}</p>

Table 3. Cont.

Epistemic Ideal Code	Code Description ¹	Example ²
Internal Structure of the Explanation		
Comprehensive	The explanation is not too simple; it is sufficiently complex.	Thin Films
		<p>Aria: So what's our reasoning? How are we . . . ? Isabella: I said that . . . Well, I said that it's true, but it doesn't really explain why, like, the appearance of the light, do you know what I mean? Like, it, it doesn't really explain why multiple colors appear. It just explains—it just—{<i>EI: Comprehensive</i>}</p>
Logically sound	The components of an explanation make reasonable, non-contradictory connections among each other.	Airbags
		<p>Chloe: Do you guys think it is safe to use sodium azide and potassium nitrate and silicone oxide in the gas generator mixtures, and sodium . . . ? Wait, so sodium . . . Sodium azide is toxic. Emma: Probably not. Isabella: I know that's, that's what I—that's one of my questions where I don't understand how that would be safe. {<i>EI: Logically sound</i>} <EI starts> Mia: Yeah. Isabella: Like, if airbags are supposed to protect you and save you, then why are they putting chemicals in it? <EI ends> Emma: And then, like, the toxic substance that, like, converts into glass . . . Isabella: Like, like, I don't understand how that came about for a safety device to have glass in it and a toxic substance that you know, converts to that.</p>
Good Communication		
Precise wording	The language used in the explanation is specific and accurate.	Hot Packs
		<p>Aria: That's not technically—{<i>EI: Precise wording</i>} <EI starts> Emma: But using this— Aria: —phrased right. Yeah, yeah, I know what you mean, but it's—it should say activation energy is the energy nee—well, like, it—you have to overcome the activation energy— Emma: To be able to start. Aria: Yes. It's not needed. Like, the activation energy is just, um, an amount. __: OK. Aria: Does that make any sense? <EI ends></p>
Clearly understandable	The explanation is well-written and easily interpreted.	Airbags
		<p>Chloe: I would agree [that model 2 is valid], um, that I just think it had kind of better language. {<i>EI: Clearly understandable</i>}</p> <p>Aria: The wording's the same, they just have different things, like, combined. Chloe: But I think the best thing was it said the airbag inflates in order to provide a force in the equal and opposite direction of the driver's motion and (inaudible) like that, like—made it better.</p>

Table 3. Cont.

Epistemic Ideal Code	Code Description ¹	Example ²
Empirical Evidence		
Coherence with empirical evidence (personally collected)	The explanation is not contradicted by empirical evidence (i.e., data that were collected systematically using scientific practices) that the students have personally gathered while assessing claims.	Thin Films Chloe: When the nail polish hit the water, it, like, spread—like, it was like a drop, but then it, like, spread and [the model] doesn't do that. {EI: Empirical evidence}
Evidentiary Support		
Evidentiary support	The explanation is supported by reasons and/or evidence.	Thin Films Grace: What do you have written down as your reasoning and evidence, Student 1? {EI: Evidentiary support}

Note. 1. The term “explanation” is used throughout the definitions in this table for consistency. However, each epistemic ideal could be applied to any epistemic product (e.g., model, argument, claim). 2. In the examples, EC codes are noted in italics within {}; duration of codes that span multiple turns is indicated within <>; EI: Epistemic ideals; RP: Reliable process. 3. Student names are pseudonyms.

Table 4. Descriptions of Reliable Epistemic Processes Codes with Examples.

Reliable Epistemic Process Code	Description	Example
Experimentation	Controlled testing of different options.	Thin Film Emma: I said, why do you think it only takes the one drop? Like, what would happen if we added more? Grace: Well, I did. Mia: Yeah, we added more. Emma: What happened? Grace: Well, I feel like after you add too many they start, like, turning into little blobs, and they sink to the bottom. That's what happens after a while. But if you have, like, two, they just kind of go inside each other, and it's just magical. Then after a while they form, like, little, tiny kind of teardrops, and they just fall to the bottom. I don't know why. {RP: Experimentation, enacted}
		Airbags Chloe: I would like to see it in real life. I think it would be better than seeing it on tel—like, on a screen is to see an airbag deployed in real life, because I've never been in an accident, thank goodness, to see that, so I think that would help better for the demonstration, for the models, have a better understanding to see it. {RP: Observation, metacognitive}
Observation	Examination of the world in real-time through human sensory perception.	Nuclear Fission Chloe: But if you, like, fly away, how would you know like when [a nuclear bomb] was supposed to hit the ground to press [a button to activate it]? Emma: So maybe you have to—at least you can see it, you can estimate, like, you can—like in physics, the freefall problems, estimate with the gravity, (laughs) how long it would take. {RP: Physics formula, metacognitive}
Physics formula	Use of a known physics formula.	

Table 4. Cont.

Reliable Epistemic Process Code	Description	Example
Process of elimination	Among a number of proposed claims, find reasons to reject each claim until one claim remains. Evidentiary support is not provided for the final remaining claim.	Nuclear Fission Chloe: Why'd we pick model two? Emma: Well, compared to the other ones, it says, "The strong nuclear force overpowered the electric static forces." And only number four also says that. The rest of them are backwards. {RP: <i>Process of elimination, enacted</i> } <RP starts> Chloe: I agree, and that refers to the text where it told us that strong, uh, nuclear forces would overpower the, uh, electrostatic forces. Grace: So then you would be able to narrow it down to two and four, and then it would be two because it says, like, the resulting nuclei will have an increased binding energy and be more stable. <RP ends>
		Hot Packs Aria: I feel like it's going to release the same amount, but, like, it's going to release it over longer period of time. If you think about, like, um, imagining you're holding a really large disc, just like that—and if you click it, like, slowly, and you kind of hold it, and then it goes over, like, to that curve slowly, so it's giving off the same amount of energy, but over a longer period of time. So it's not giving much at a time. So if you just click, it just flips, but if you click it slowly, it doesn't give as much energy at a time." {RP: <i>Thought experiment, enacted</i> }
Thought experiment	Working logically through an imagined scenario.	

2.4.2. Quality Talk Coding

Two trained researchers independently coded the discussions and came together to reconcile any disagreements in accordance with the Quality Talk Coding Manual [72]. In order to facilitate analysis, we segmented the discussion into episodes of talk based on the authentic question events [73]. Each authentic question event began with an authentic question asked by a student and included all responses generated by students in response to that question. Responses related to the authentic questions were coded for individually (i.e., elaborated explanation, EE) and collectively constructed argumentation (i.e., exploratory talk, ET; cumulative talk, CT). See Table 5 for code descriptions and examples.

2.5. Data Analysis Plan

Of the six cycles of QT_S discussions, four were selected for analysis: two from fall and two from spring (see Table 1). The lesson on Soap Bubbles was not analyzed due to technical malfunctions with the camera, and the lesson on Tesla Coil was not analyzed due to a new student joining the group. Video and audio data from these QT_S small-group discussions were transcribed by a professional transcriber into word processing documents. These transcription files were uploaded to a qualitative data analysis software (ATLAS.ti 7) to facilitate coding and analysis.

For RQ 1, we identified the categories of epistemic cognition and argumentation invoked during the QT_S discussions via qualitative coding and through iterative coding and reconciling by two raters. For RQ 2, we detected the differences in students' epistemic cognition and argumentation due to contextual factors by examining the frequency of the epistemic cognition and argumentation codes during the model-based portion of the QT_S discussions. For RQ 3, we investigated how students' epistemic cognition related to their authentic questioning and argumentation by examining the extent to which epistemic cognition codes, the authentic question code, and argumentation codes co-occurred or were closely related to one another (e.g., where one code tended to immediately follow another code) and then checking the transcripts to verify patterns.

Table 5. Descriptions of Quality Talk Codes with Examples.

Discourse Code	Short Code	Description	Examples
Authentic Questions	AQ	Question in which more than one acceptable answer is possible, the speaker genuinely is interested in the responses of others, and there is no known “correct” answer.	Thin Films Grace: What causes the nail polish on the surface to, like, spread out? Do you guys know? {AQ}
			Airbags Chloe: I think, going back to Emma’s question, it’s kinda just, I feel it saves you to an extent, but I think without your seatbelt, like, if you don’t wear your seatbelt and the airbag’d deployed I feel like it wouldn’t really do much, but I feel like the seatbelt keeps you grounded. Like, I think in the demonstration they were, like wearing, like a lap belt, and, like, when they, like, came up, like, the airbag hit them and they went back down. But, like, without the seatbelt they would probably go over the airbag, you know what I mean—and then, like, crash. {EE}
Elaborated Explanation	EE	Response to an AQ where an individual offers an explanation that includes a claim with multiple pieces of reasoning and/or evidence.	Thin Films Chloe: Oh, wait, wait, wait, I have a—I have a answer for (3)’s question. Because technically these lights aren’t truly white light. They’re not purely white, so they might not fully have all the colors. {ET} <ET starts> Grace: Oh, but then whenever you look at the outer edge it has (inaudible). Oh, I see some blue in there when you go like this. Mia: Wait, but— Isabella: Like, in between the green, but it still goes from green to pink. Mia: Wait, but [didn’t] they reflect— Isabella: (inaudible). Mia:—all the light (overlapping dialogue; inaudible)? Isabella: You have to, like, hold it at an angle. Aria: Oh, yeah. Emma: But (teacher) even said there were, like, close, but they weren’t, like . . . Chloe: I would expect (overlapping dialogue; inaudible). Grace: (overlapping dialogue; inaudible) Isabella: Wait, guys, I can see different colors now. Chloe: I don’t think [they’re] (inaudible). Grace: Oh, it’s all about that angle. Mia: I think (inaudible). Chloe: It is. Like, look at it on an angle. Then you see, like, all different colors (inaudible). It’s like now I see blue and then I see pink. Emma: So it has to do with the angle— Grace: Oh, yeah. Emma:—that you’re looking at it. <ET ends>
Exploratory Talk	ET	Collaborative exchange where multiple students build on and share knowledge, evaluate evidence, or weigh different options over multiple turns. ETs are differentiated from CT by the presence of a challenge.	Thin Films Mia: Yeah, wavelength dictates color. {CT} <CT starts> Aria: I feel like each color, like, each different— Isabella: Oh, yeah. Aria:—spot of color on the piece, piece of film is giving a different wavelength, so our eyes are perceiving, like, this red or purple or whatever color. Grace: Yeah, it has to (inaudible)— Aria: But it doesn’t mean they’re bouncing off of each other. They’re just— Grace: Yeah, it has to do with the wavelength. Aria: Yeah. Grace: Whatever she said, bouncing off of each other. Isabella: I mean, it’s, like, implying that all wavelengths are the same, like, all colors have the same wavelength. Grace: But they just bounce off each other. Isabella: But they just bounce off each other, like . . . Aria: I don’t think that’s true. Isabella: Yeah. Grace: Yeah. Nice try. <CT ends>
Cumulative Talk	CT	Collaborative exchange where multiple students build on and share knowledge in a way that is positive, but not critical, over several turns. CTs do not contain the element of challenge that characterizes ETs.	

Note. QT codes are noted in bold within {}.

3. Results

3.1. RQ1. Epistemic Ideals, Reliable Processes, and Argumentation Invoked in Science Discussions

3.1.1. Epistemic Ideals

Our qualitative open-coding procedure resulted in a set of 10 codes related to students' epistemic ideals. Then, these codes were organized into five epistemic ideal categories: connections to other knowledge, internal structure of the explanation, good communication, empirical evidence, and evidentiary support (see details in Table 3). With respect to the frequency of these epistemic ideals, there was a notably wide variation between the different categories. As shown in Table 6, students most frequently made 'connections to other knowledge,' as evidenced through 49 instances. In contrast, there was only one instance of the category 'empirical evidence,' as represented through the code 'coherence with personally collected empirical evidence.' In the following descriptions of each category, we refer to the examples provided in Table 3.

Table 6. Frequency Table of EI Codes Across Discussions.

Epistemic Ideal Code	Fall				Spring			
	D1. Airbags		D2. Nuclear Fission		D3. Thin Films		D4. Hot Packs	
	Model-Based	Open-Ended	Model-Based	Open-Ended	Model-Based	Open-Ended	Model-Based	Open-Ended
Connections to other knowledge (total count = 49)								
Coherence with normative disciplinary knowledge	2	2	3	8	6	1	5	5
Coherence with personal experience	0	1	0	0	0	0	0	7
Coherence with the personal experience of a layperson	0	1	0	0	0	0	0	0
Coherence with prior knowledge (other)	0	2	0	2	1	3	0	0
Internal structure of the explanation (total count = 7)								
Comprehensive	0	2	0	0	1	2	0	0
Logically sound	0	1	0	0	1	0	0	0
Good communication (total count = 2)								
Precise wording	0	0	0	0	0	0	1	0
Clearly understandable	1	0	0	0	0	0	0	0
Empirical evidence (total count = 1)								
Coherence with empirical evidence (personally collected)	0	0	0	0	0	1	0	0
Evidentiary support (total count = 3)								
Evidentiary support	0	0	0	0	1	0	2	0

The vast majority of identified instances of epistemic ideals related to how epistemic products connected to other knowledge (e.g., personal experience or prior knowledge). Within this category, the most commonly invoked epistemic ideal was noted by instances coded as 'coherence with normative disciplinary knowledge' (i.e., coherence with NDK). As evidenced by the 32 instances of this code, this ideal generally involved students expressing normative science knowledge that related to what they learned from the QT₅ science lesson, such as the scientific articles. As shown in the example in Table 3, Aria (student names are pseudonyms) explicitly referred to an article assigned from the nuclear fission lesson in response to Emma's authentic question. Likewise, in the airbag discussion, Chloe made a connection to their shared prior knowledge from the demonstration video that they watched together in class.

With respect to the category of 'internal structure of the explanation,' students evaluated whether explanations were sufficiently complex (i.e., comprehensive) or internally consistent (i.e., logically sound). In alignment with the two codes in this category, students expressed hesitations about whether to accept a claim or explanation because they felt it was incomplete or missing important explanatory components or they speculated about

the logic in the explanations. For example, when discussing the use of the chemical sodium azide in airbags, students wrestled with whether the use of the highly toxic sodium azide was contradictory to the use of an airbag as a safety device. That is, students stated the ideal of being logically sound was not sufficiently met.

Instances associated with the ‘good communication’ category pertained to epistemic ideals related to the language and comprehensibility of explanations. The codes that made up this category included ‘precise wording’ and ‘clearly understandable’ and were notably infrequent in the discussions, each occurring only once. In the example, Chloe expressed that they accepted Model 2 because of the language, and then went on to describe the exact phrasing of the explanation that they were referring to as it related to this ideal. These statements exemplified how the student accepted the claim on the basis of its clear, understandable language in comparison to the other claims.

The category of ‘empirical evidence’ related to the epistemic ideal that students used to seek coherence with empirical evidence that was personally collected. However, formal data collection was not a component of the science lessons in the present study, and students only conducted hands-on experiments during some lessons. Despite this, there was one instance where a student made a connection with empirical evidence. In the thin films lesson, students had an opportunity to engage in a hands-on activity. They submerged a scrap of black construction paper in water, added a drop of nail polish onto the water, and then observed how the nail polish formed a layer on the surface of the water as well as how the paper looked once it was lifted out of the water. Thus, during the discussion about thin films in the example, Chloe referred to the empirical evidence and emphasized that the model did not seem to align with what was observed during the demonstration.

Finally, we identified one code that did not fit within any of the five proposed by Chinn and colleagues [30]. We termed that code and the broader category ‘evidentiary support.’ As illustrated by the evidentiary support code, students either accepted a claim because it was supported by reasons and/or evidence or they held their peers accountable for providing reasoning and/or evidence. Notably, this was different from empirical evidence where students provided or referred to empirical evidence that was personally collected to support a claim. An example would be Grace explicitly prompting their group for reasoning and evidence to help evaluate a claim.

3.1.2. Reliable Epistemic Processes

Through our open-coding process, we coded five types of reliable epistemic processes: experimentation, observation, physics formula, thought experiment, and process of elimination (see code descriptions and examples in Table 4). Compared to the frequency of epistemic ideals, there were fewer instances of reliable epistemic processes, which added up to 11 instances across four QT_S discussions (see Table 7).

Table 7. Frequency Table of Reliable Processes Codes Across Discussions.

Reliable Epistemic Process Code	Fall				Spring			
	D1. Airbags		D2. Nuclear Fission		D3. Thin Films		D4. Hot Packs	
	Model-Based	Open-Ended	Model-Based	Open-Ended	Model-Based	Open-Ended	Model-Based	Open-Ended
Experimentation	0	1	0	0	0	1	0	0
Observation	0	1	0	0	0	1	0	3
Physics formula	0	0	0	1	0	0	0	0
Process of elimination	1	0	1	0	0	0	0	0
Thought experiment	0	0	0	0	0	0	0	1

Each of these five codes involved a different method used to construct epistemic products, that is, a process used to establish knowledge, models, explanations, or theories.

For instance, ‘experimentation’ referred to using controlled testing as a reliable process to produce an epistemic product. Similarly, for other reliable process codes, students considered examining the world through human sensory perception (i.e., observation), referring to a known physics formula (i.e., physics formula), logically thinking through an imagined situation (i.e., thought experiment), or applying a process of elimination as reliable processes to obtain an epistemic product. For instance, during the discussion on thin films, Grace referred to her experimentation as a reliable process to explain what happened after adding one more drop of nail polish on the water surface. An example of a thought experiment as a reliable process was identified in the discussion on hot packs as shown in Table 4. Aria was thinking through an imagined scenario where the energy would be released over a longer period of time when one clicked the disc more slowly.

As we attempted to identify the categories of reliable processes in student discourse, we also noted the ways in which students used the schemas that specified these reliable processes (i.e., enacted, evaluative, and metacognitive). An example of students enacting a reliable process would be picking a model that they deemed to be correct through the process of elimination (see Table 4). Specifically, when students discussed which model to pick for the nuclear fission discussion, they did not argue why the selected Model 2 was correct. Instead, Emma and Grace eliminated other models because one statement in the rest of the models did not seem to align with the information provided in the reading. As noted earlier in the qualitative coding for reliable processes, even though it was not viable for participants to enact all possible reliable processes, students could still speculate about the reliable processes by evaluating and metacognitively thinking and talking about them. As a case in point, in the discussion on airbags, Chloe expressed a metacognitive belief regarding observation as a reliable process, stating that having first-hand observations from a real-life experiment would help the group better understand the models than a video demonstration. Similarly, in the discussion on nuclear fission, Emma expressed her metacognitive belief about using a physics formula as a reliable process that could estimate how long it would take a nuclear boom to hit the ground.

3.1.3. Argumentation

The final piece related to the first research question involved students’ use of argumentation in the discussion. Our argumentation coding was operationalized through the response codes of the Quality Talk coding. Three argumentation codes were used to identify episodes of talk that evidenced both individually (i.e., elaborated explanation, EE) or collectively constructed argumentation (i.e., cumulative talk, CT; exploratory talk, ET). The frequency of each argumentation code was generally balanced across the four QT_S discussions (see Table 8), but EEs ($n = 41$) occurred more frequently than ETs ($n = 10$) or CTs ($n = 24$).

Table 8. Frequency Table of Quality Talk Codes Across Discussions.

QT Code	Fall				Spring			
	D1. Airbags		D2. Nuclear Fission		D3. Thin Films		D4. Hot Packs	
	Model-Based	Open-Ended	Model-Based	Open-Ended	Model-Based	Open-Ended	Model-Based	Open-Ended
AQ	2	11	1	17	13	18	4	15
TQ	0	1	0	1	1	1	3	1
EE	1	7	2	9	6	6	4	6
ET	0	2	1	1	3	2	1	0
CT	2	4	1	5	3	4	1	4

EEs are individual explanations (i.e., an uninterrupted turn by a single speaker) that include a claim and multiple pieces of reasoning and/or evidence. For instance, in the example shown in Table 5, Chloe first started with a claim regarding the necessity of a

seatbelt in response to a question about the pros and cons of airbag. Following this claim, they provided evidence or reasoning, one being the demonstration they watched in class and the other being the reasoning derived from the demonstration.

The code for ET captures episodes of collaborative, group-constructed discourse where students weigh different arguments over multiple turns and is characterized by the use of a challenge. For example, in the discussion on thin films (see Table 5), Grace challenged Chloe's reasoning about why the colors went from green to pink on the thin film but not purple or blue. Chloe first proposed that it was because the lights were not truly white light. However, Grace challenged this claim by referring to their observation of the thin film when the black paper was held at different angles and argued that the reason could be the angle of perception.

In contrast, the code for CT captures episodes of talk that are collaborative, group-constructed exchanges where multiple students build knowledge but not critical way. That is, there is no presence of a challenge in a cumulative talk episode. According to the example in Table 5, four students co-constructed their understanding of the relationship between wavelength and colors seen on the thin film without challenging each other. Together they built their evaluation of a claim in the presented model and concluded that the claim was not correct.

3.2. RQ2. Contextual Factors Related to Students' Epistemic Cognition and Argumentation

Scholars have increasingly acknowledged the influence of context, including factors such as domain- and task-specificity of phenomena, on epistemic cognition [10,32]. In order to explore our second research question, the authors met to identify trends with regard to students' epistemic cognition and argumentation as related to change in the contextual factors from fall to spring. Specifically, the contextual factors of interest included the model format (i.e., the scientific model task shifted from selecting one best model from four models to evaluating three separate claims) and explicit instruction provided to the students (i.e., the focus of QT_S discourse lessons and QT_S catalyst shifted from authentic questions to argumentation components). Herein, we present three trends that demonstrate changes in students' epistemic ideals, reliable processes, and argumentation from fall (i.e., discussions on airbags and nuclear fission) to spring (i.e., discussions on thin films and hot packs) in the model-based portion of the discussion, as it was the portion of the discussion that was more sensitive to changes related to model format.

3.2.1. Students Evidenced Increased Use of Epistemic Ideals

We identified two primary trends regarding shifts in the epistemic cognition codes for the model-based portion between fall and spring. First, with regard to epistemic ideals, there were substantially more occurrences of coherence with normative disciplinary knowledge ($n = 11$) and evidentiary support ($n = 3$) in spring compared to fall ($n = 5$ and $n = 0$, respectively; see Table 6).

As illustrated in Excerpts 1 and 2, during both spring discussions (i.e., thin film and hot packs discussions), students invoked the standard that acceptable claims must be supported by reasons and evidence (i.e., evidentiary support). They systematically evaluated each of the three claims presented. In Excerpt 1, Grace asserted that the first claim was correct because it was the only claim with evidence, meaning that this claim met a necessary criterion (i.e., claims must be supported by evidence). Then, Aria endorsed this epistemic ideal and prompted the group to provide evidence by asking, "What's your evidence behind it?"

Students also referred to the revised QT_S catalyst that was focused on the key argumentation components (i.e., claim, reasoning, and evidence) to probe for evidence of each claim in the provided scientific model. Similar to the question that Aria asked in Excerpt 1, in Excerpt 2 from the thin films discussion, Grace asked, "What do you have written down as your reasoning and evidence?" (Note: In each excerpt, EC codes are noted in italics

within {}; QT codes are noted in bold within {}; duration of codes that span multiple turns is indicated within <>. EI: Epistemic ideals; RP: Reliable process.)

Excerpt 1: Hot Packs

Grace: I think that's the only [claim] that has evidence. {EI: Evidentiary Support}

Aria: Yeah, same. So what's your evidence behind it?

Isabella: I talked about this little—this little [doodah], this—graph, about the activation energy. I said that by clicking it—that it provides the activation energy for the reaction to start occurring. {EI: Coherence with NDK}

Excerpt 2: Thin Films

Grace: When a drop of nail polish is dropped onto a warm surface, the lower density of the nail polish and the molecular attraction of the molecules prevent it from mixing with the water? Now, what does everyone think about this?

Chloe: I think it's false.

Isabella: I think it's true.

Grace: What do you have written down as your reasoning and evidence? {EI: Evidentiary support}

Aria: Yeah, why do you think it's false?

These excerpts are examples of how both the changing context of the model format and the explicit instruction affected the ways students evidenced epistemic cognition via argumentation. The structure of the model and lessons on argumentation made it more likely that students would surface their epistemic cognition (i.e., epistemic ideals), which in turn, made their epistemic cognition public and available for scrutiny by their peers via argumentation. The process of argumentation, a scientific epistemic practice, could then lead to improvements in epistemic cognition.

3.2.2. Students Evidenced Decreased Use of Process of Elimination

The second trend regarding the change in students' epistemic cognition pertained to students' use of reliable processes for the model-based portion between fall and spring. There were only two instances of reliable processes in the model-based portions of the discussions, and notably, these both occurred during the fall discussions. That is, students used the process of elimination in both airbags and nuclear fission discussions to identify the normative model that they held to be true. Specifically, students used the process of elimination to identify the model they believed was the most appropriate without challenging each other, probing for alternative arguments, or requesting further justification for the claim. In essence, in fall, students engaged in a process of elimination by narrowing the options provided in the model, a process they considered reliable for achieving their epistemic end (see Excerpt 3). In contrast, after the change in the model format, students did not use process of elimination when discussing the provided scientific model in spring.

Excerpt 3: Nuclear Fission

Chloe: Why'd we pick model two?

Emma: Well, compared to the other ones, it says, "The strong nuclear force overpowered the electric static forces." And only number four also says that. The rest of them are backwards. {RP: Process of elimination, enacted} <RP starts> {EE}

Chloe: I agree, and that refers to the text where it told us that strong, uh, nuclear forces would overpower the, uh, electrostatic forces.

Grace: So then you would be able to narrow it down to two and four, and then it would be two because it says, like, the resulting nuclei will have an increased binding energy and be more stable. <RP ends>

The process of elimination is not a typical, normative scientific practice. Again, this change in the use of reliable processes was likely due, in part, to the change in model format but also likely resulted from QT instruction in argumentation, which emphasized more normative epistemic practices in science than the process of elimination.

3.2.3. Students Evidenced Increases in EE, ET, and CT

The last trend focused on changes in students' argumentation as evidenced by the individual- and group-constructed argumentative responses. Notably, there were more instances of EE ($n = 10$), ET ($n = 4$), and CT ($n = 4$) in spring than fall ($n = 3, 1,$ and 3 , respectively) during the model-based portion of the discussions.

The increased occurrences of these argumentation codes indicated an improvement in the quality of student argumentation in general, but as we examined students' EEs in the transcripts, we also noted that the EE generated by Emma in spring appeared to have a higher quality than the EE she initiated in fall. For instance, in Excerpt 3, students were discussing why they would pick Model 2 from the four models presented to them during the nuclear fission discussion. In response to this question, Emma generated an EE that indicated a process of elimination, a non-normative reliable process. That is, as long as a claim includes a statement that the students think is wrong or is reversed from the statement that they hold to be true, it is automatically eliminated regardless of what remains in the claim. Such reasoning did not directly explain if the model was scientifically acceptable or not. By contrast, in spring, as students were evaluating the second claim in the provided model about hot packs, Emma initiated an EE which included an explicit reference to the evidence that was closely related to the claim. As shown in Excerpt 4 below, Emma used scientific evidence to explain why the second claim was considered wrong.

Excerpt 4: Hot Packs

Emma: Yeah, whenever—in this article it says—like, it's talking about entropy, it says, like, an increase in entropy is represented by a positive value for delta S, which is, like, an endothermic reaction. So that's kind of like . . . But this is saying, like, an increase in entropy makes it an exothermic reaction. So it's kind of, like, saying the opposite in here. {EE}

Aria: Yeah.

Emma: That's what I used for my evidence, like, down here—

Such discourse is reflective of more normative argumentation processes in science, where counterclaims and rebuttals must be supported with reasoning and evidence. As students evaluated each separate claim, they needed to provide reasoning and evidence to support a correct claim as well as to refute an incorrect claim. The ability to evaluate claims via argumentation was likely strengthened through explicit instruction on argumentation.

3.3. RQ3. The Relation of Epistemic Cognition to Authentic Questioning and Argumentation

For our third and final research question, we looked at how students' epistemic cognition related to their authentic questioning and argumentation. Specifically, we produced a set of figures to demonstrate the timelines and the codes of epistemic cognition, argumentation, and authentic questions across the entirety of four QT_S discussions, which we examined in combination with the transcripts to synthesize trends (see Figures 4–7). In sum, we identified two major trends: (a) the use of epistemic ideals was associated with the initiation of authentic questions, and (b) argumentation involved the use of epistemic ideals.

3.3.1. The Use of Epistemic Ideals Was Associated with the Initiation of Authentic Questions

As shown in Figure 4 along with the transcripts, a pattern emerged whereby (a) the use of epistemic ideals seemed to trigger the initiation of an authentic question and (b) the initiation of an authentic question seemed to lead to the use of epistemic ideals. For the first

trend, we explored the transcripts where the use of an epistemic ideal co-occurred with or immediately preceded the initiation of an authentic question to identify the relationship. For the second trend, we explored the transcripts where the initiation of an authentic question preceded the use of an epistemic ideal to verify the finding.

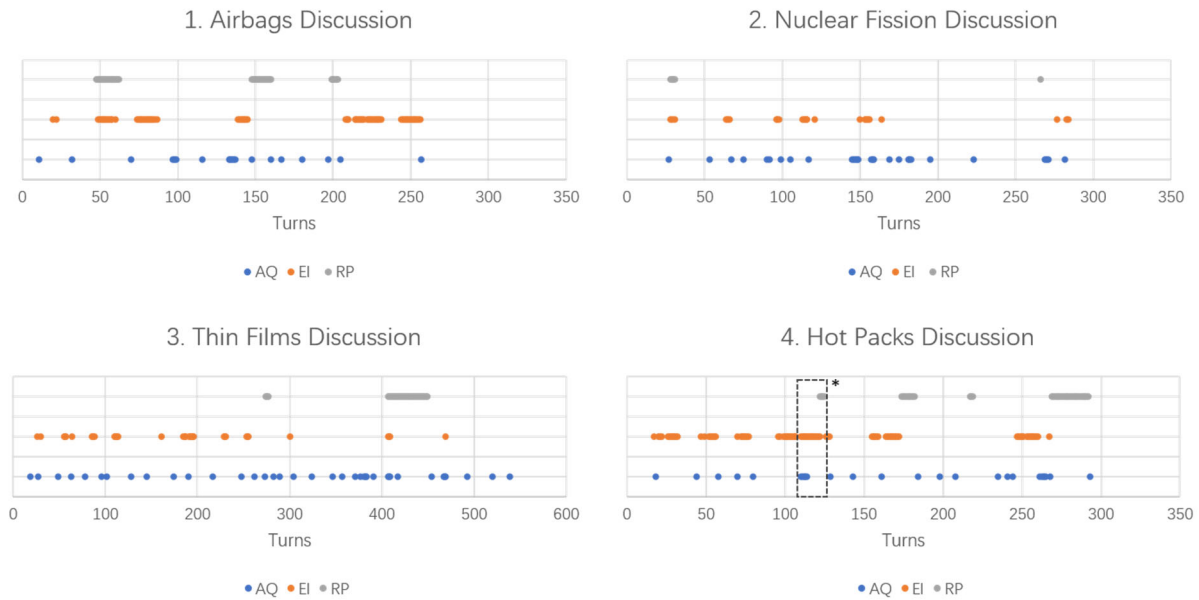


Figure 4. Occurrences of Authentic Questions, Epistemic Ideals, and Reliable Processes in Four Discussions. Note. 1. AQ = Authentic Questions, EI = Epistemic Ideals, RP = Reliable Processes; * Episode of talk exemplified in Excerpt 5. 2. Please also note that the thin films discussion had more turns than the other three so the scale for the X-axis is different from the rest.

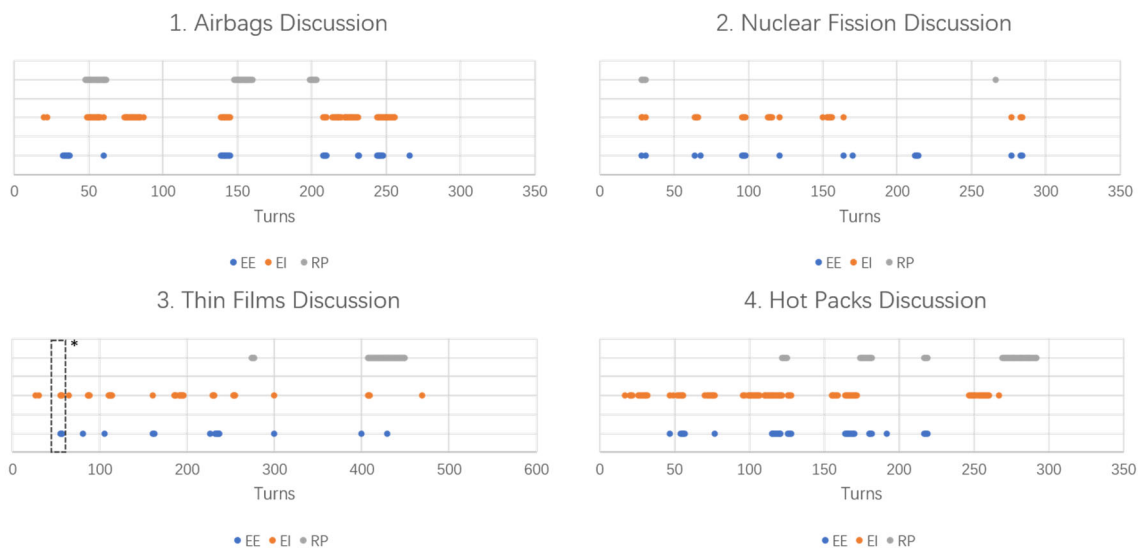


Figure 5. Occurrences of Elaborated Explanation, Epistemic Ideals, and Reliable Processes in Four Discussions. Note. EE = Elaborated Explanation; * Episode of talk exemplified in Excerpt 6.

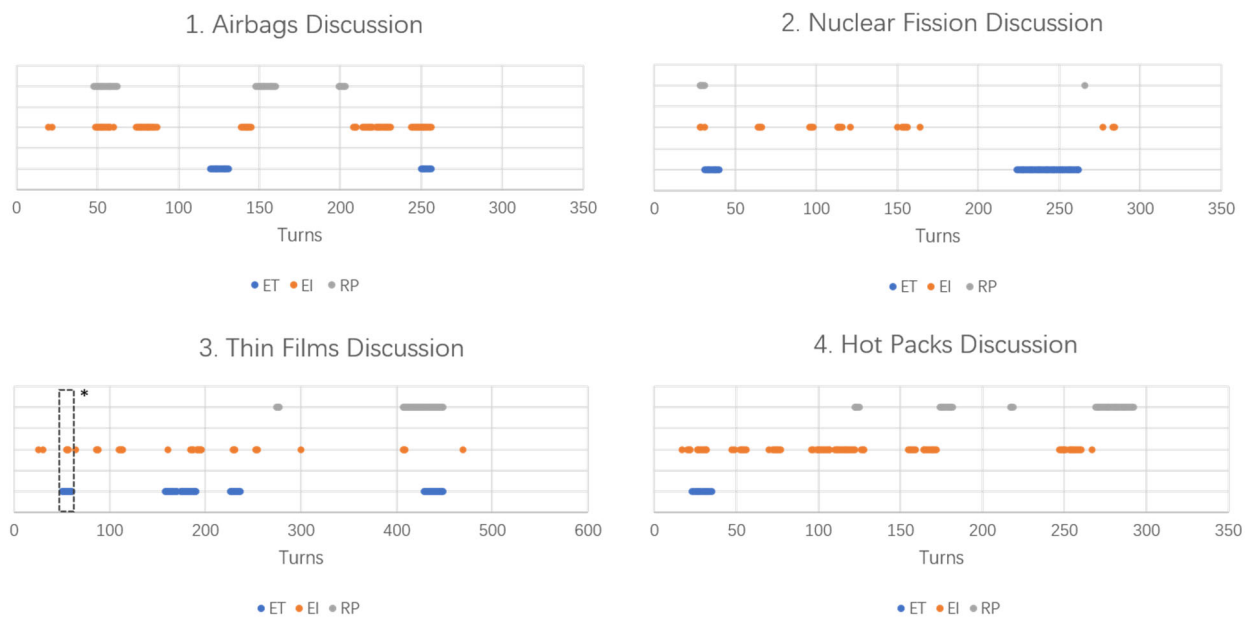


Figure 6. Occurrences of Exploratory Talk, Epistemic Ideals, and Reliable Processes in Four Discussions. Note. ET = Exploratory Talk; * Episode of talk exemplified in Excerpt 6.

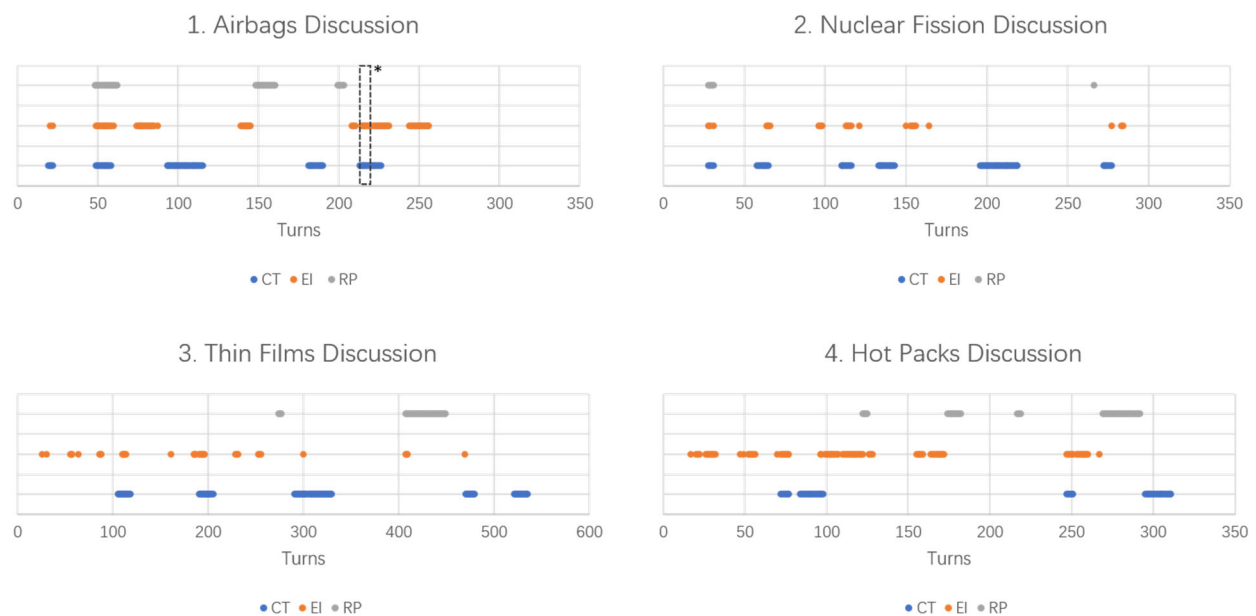


Figure 7. Occurrences of Cumulative Talk, Epistemic Ideals, and Reliable Processes in Four Discussions. Note. CT = Cumulative Talk; * Episode of talk exemplified in Excerpt 7.

An example of the first trend is evidenced in Excerpt 5, which is denoted in a block with an asterisk in Figure 4. At the beginning of this excerpt, Isabella’s turn was coded both as an authentic question and an epistemic ideal, that is, the two codes co-occurred. In this turn, Isabella first connected their everyday experience of boiling chocolate with the changing state of hot packs. The idea that a reusable hot pack, after being boiled, did not turn solid again at room temperature did not seem to cohere with their personal experience about the substance of chocolate. Thus, students had to reconcile the lack of coherence between the two situations to fully understand the phenomenon under which sodium acetate in hot packs behaved differently than chocolate. In this scenario, the epistemic ideal of coherence with personal experience was not met and thus prompted a productive authentic question asked by Isabella, approaching the end of the same turn,

about why a hot pack did not get solid again at room temperature like chocolate. In this excerpt, the scrutiny of epistemic ideals facilitated productive discussions around students' misconceptions through the initiation of an authentic question.

Excerpt 5: Hot Packs

Isabella: OK, and then, like, what you were saying before, how you boiled it—and then it stays a solid—I mean, a liquid—like, that's, like, different. You know, like, when you heat chocolate, OK—it's, like, solid at first, and then you heat it, and then it turns into, like, a b—then it turns back into a solid. So, like, why wouldn't this do that? <EI: Coherence with PE> {AQ}

Aria: I think the freezing points are different. Like, with ice, if you melt it, it's just gonna stay, like, water, unless you put it back into the fridge again—<EI: Coherence with NDK> {EE}

Isabella: Oh.

Aria: —because the freezing point is all (inaudible).

Isabella: OK.

Aria: Anyway, um, but with chocolate, it probably has a really high freezing point—

Isabella: It's (inaudible)—

Aria: —'cause at room temperature it's a solid, right?

Isabella: Yeah. But it's weird, because this—like, at room temperature it can be a solid or a liquid. {RP: Observation} {EI: Coherence with PE}

Aria: Yeah, that's kind of weird, huh?

Grace: Yeah, 'cause yours is . . . Well, mine's as hard as can be, and hers is like a gel.

Isabella: Or, like, they're just doing it right now, like, they just boiled theirs, and it's, it's gonna stay a solid. Like, it's not gonna go back to . . . I mean, it's gonna stay a liquid. It's not gonna go back to a solid.

Aria: Yeah, I don't know how they engineered it to—so that it stays . . . I feel like it's probably the chemical properties, because it says, "But it can exist as, as a liquid at a much lower temperature," like, lower than the freezing point, "and it's extremely stable." I don't know why that is, but I think—{EI: Coherence with NDK} {EE}

The use of epistemic ideals seems to trigger the initiation of authentic questions, but authentic questions also probe for the use of epistemic ideals in student responses to these authentic questions. Again, as shown in Excerpt 5, Isabella's authentic question led to multiple student responses that connected to the normative disciplinary knowledge (by Aria) as well as personal experience (by Isabella). For example, in response to Isabella's authentic question, Aria was seeking coherence with the normative disciplinary knowledge by citing a piece of evidence from the student's prior knowledge about the melting of ice and later brought up a reference reading regarding the chemical properties of hot packs.

3.3.2. Argumentation Involves the Use of Epistemic Ideals

The second trend pertained to the co-occurrence of argumentation and epistemic ideals, indicating frequent use of epistemic ideals in constructing argumentation. Specifically, EE co-occurred with epistemic ideals approximately 50% of the time (see Figure 5). Further, as shown in Figures 6 and 7, ET and CT also co-occurred with epistemic ideals approximately 60% of the time. The excerpts presented in the sections below provide additional evidence to bolster our argument regarding this trend.

Students' EE involved the practice of epistemic ideals. For example, in Excerpt 6 (also see the block noted in Figure 5), students were evaluating one claim in the student

model during the thin films discussion, that is, “When a drop of nail polish is dropped onto a water surface, the lower density of the nail polish and the molecular attraction of the molecules prevent it from mixing with the water.” During the discussion, Aria generated an EE by utilizing normative disciplinary knowledge and referred to the bonding between molecules of the nail polish and the bonding between water molecules as her reasoning.

Excerpt 6: Thin Films

Isabella: Why else do you think the, the first one [claim is false]—{AQ}

Chloe: No, I just—it doesn’t make sense that the molecular attraction of the molecule, uh, (inaudible)—{EI: *Logically sound*} {ET} <ET starts>

Aria: OK, think of it this way: you have, um, like, this drop of nail polish in just, like, water. Where was the water? Was it on the film or something?

Emma: It was in, like, a little plastic tub.

Isabella: It was (inaudible).

Emma: Yeah.

Aria: OK. So you have the nail polish, and the nail polish molecules attract one another, so they want to stick, stick together, sort of like . . . And the water has hydrogen bonding, your favorite type of bonding, right? And then they want to stay together, so the nail polish is not gonna just mix with the water, because they’re still, like, together, because (inaudible) molecular forces are bonding them together, that they’re not separated. Does that make sense? {EI: *Coherence with NDK*} {EE}

Grace: Basically, the water molecules don’t want to get a divorce, and the nail polish ones don’t either, so they just kind of—

Aria: Yeah.

Grace: —coexist. <ET ends>

In addition, argumentation was found to co-occur with epistemic ideals as evidenced through ETs and CTs. As illustrated in Excerpt 6, an ET occurred following an authentic question. In this example, some students stated the first claim in the student model was true, whereas others disagreed. Chloe struggled to accept the first claim. The statement that molecular attraction prevents the nail polish from mixing with water did not appear to be logically sound to her. The use of this epistemic ideal (i.e., logically sound) led to Chloe’s challenge in the group, which characterized this ET. In response to Chloe’s challenge, Aria provided an EE to demonstrate that this claim cohered with normative disciplinary knowledge as explained earlier. After this EE, the group collectively decided that the first claim was valid and moved onto the next claim in the model. Collaboratively, students used various epistemic ideals to construct argumentation as a group by raising a challenge and responding to the challenge.

During episodes of talk where co-constructed understandings occurred, but without challenging each other (i.e., CT), it appeared that epistemic ideals were also enacted as part of the knowledge building process (see Figure 7). For example, in Excerpt 7, students were co-constructing a response to an authentic question “Do airbags cause more injuries or prevent more injuries?” As Isabella and Mia built upon each other’s response about how airbags could prevent people from smashing into the windshield, they referred to their prior knowledge from a demonstration video that they watched in class.

Excerpt 7: Airbags

Mia: I feel like . . . I feel like a lot of airbags also do is they, like, keep you from (inaudible) if you were to smash into the windshield, too. {CT} <CT starts>

Isabella: Yeah, in the one—in the one—{EI: *Coherence with PK*} <EI starts>

Mia: Cause if you smash in the windshield . . .

Isabella: —demonstration, like, without the—without the airbag, it showed, like, the person—

Mia: Yeah.

Isabella: —it's not an actual person, but—

Mia: The (inaudible) going through.

Isabella: —but the person, like, going through the window, and, like, you could see (inaudible). <EI ends>

Mia: Yeah, (inaudible), like, if you hit it, it's gonna shatter. It's like a glass that stays together. So, like, if you go through it, it's not gonna shatter around and you're gonna be stuck in it, and you're gonna have (inaudible). <CT ends>

These excerpts are examples of how students' epistemic cognition interacted with authentic questioning and argumentation during small-group discussions, indicating a close relationship of epistemic cognition to authentic questioning and argumentation. It also enhances the understanding about how epistemic cognition can be enacted and supported by authentic questioning and argumentation during small-group, QT_S discussions in science.

4. Discussion

Modern science education standards are focused on literacy practices, including the ability to engage in scientific thinking and argumentation [4]. Society expects students to be able to readily evaluate, accept, and use scientific knowledge as they reason about science in their own lives [35,74]. Despite the strong rationale behind incorporating argumentation into science education, it remains limited in most science classrooms [40].

In response, our study contributes to science education and literacy research by exploring how small-group discussions can be used as a pedagogical tool to help students acquire the epistemic cognition and argumentation practices necessary to be thoughtful critics of scientific claims inside and outside of the classroom [10]. In prior work, we implemented QT_S, a teacher-led, small-group discussion approach designed to promote students' scientific oral and written argumentation skills and identified increases in students' scientific argumentation [27]. In the present study, we utilized the AIR model as the framework to examine and analyze high school chemistry students' epistemic cognition (i.e., epistemic ideals and reliable processes) [30] and scientific argumentation as they participated in small-group, scientific discussions about, around, and with scientific text or content.

4.1. RQ1. Documented Evidence of Students' Epistemic Cognition

4.1.1. Epistemic Ideals

Most of the identified epistemic ideals in the present study are in alignment with those in the extant literature. For instance, four of the categories we identified, namely connections to other knowledge, internal structure of an explanation, empirical evidence, and good communication, align directly with the categories of epistemic ideals (i.e., "connection to other knowledge," "internal structure of an explanation," "clearly presented and understandable," and "present and future connections to empirical evidence") set forth by Chinn and colleagues [30] (pp. 433–434). The codes organized into these four categories also appear to align with the broader extant literature. For instance, the epistemic ideal of logically sound is similar to the criteria of plausibility [69,75] or logical consistency [76]. The epistemic ideal of comprehensive aligns with what scientists consider to be good models, which can strike a balance between complexity and parsimony [77]. Similarly, clearly understandable, which is classified into the category of good communication in our study, is a communicative criterion that scientists use to evaluate models. Indeed, as Pluta et al. argued, if a successful model is not presented in a way that scientists understand, it will not be accepted and thus, the communication criterion has to be fulfilled prior to evaluating epistemic quality [70].

However, one epistemic ideal (i.e., standards of testimony) proposed by Chinn and colleagues was not identified within these four science discussions. Chinn et al. proposed a category of epistemic ideals related to standards of testimony, which indicates the criteria that must be met to believe testimony from others [30]. For example, a student could use quotes from a climate scientist to support their argument about climate change as a result of referring to the expert in the area. In this dataset, we found no instances where students referred to standards of testimony. They occasionally brought forth testimony as evidence (e.g., provided experiential accounts from their relatives) and tended to use personal experience when discussing familiar topics. However, they were stating that a knowledge claim cohered with the information they received as testimony, rather than asserting that a particular testimony was valid according to some standard. This indicates that students may need more explicit instruction that guides them to evaluate the source of evidence within the context of small-group discussions so that they may be more likely to employ the standards of testimony.

4.1.2. Reliable Processes

Throughout the discussions, we found evidence of students using a number of normative reliable processes, including experimentation and observation, which are consistent with Chinn et al. [30]. However, students also adopted non-normative processes when the situation allowed for it. For example, students used the process of elimination when they were asked to select one normative scientific model from four options for the airbags and nuclear fission discussions. It is important to note that the process of elimination is considered a non-normative process because it does not align with scientists' epistemic practices as delineated in prior research [34–36]. A plausible reason for the use of the process of elimination could be that when students were given four models with overlapping claims to choose from, when the goal of the task emphasized determining the best model, students were likely to apply a simple heuristic to narrow down the options. Another possible reason is related to the level of knowledge students had about argumentation and normative reliable processes in science. The explicit instruction on argumentation delivered in spring may have made it less likely for students to use non-normative processes such as the process of elimination when evaluating scientific claims.

4.2. RQ2. Model Format and Explicit Instruction in Relation to Epistemic Cognition and Argumentation

The epistemic practices that individuals engage in vary widely depending on contextual factors [15,78]. In the current study, the influence of context pertained to changes in the model format and explicit instruction provided during the QT₅ intervention. Following the changes in these two contextual factors, instead of using the process of elimination, students tended to use epistemic ideals such as coherence with normative disciplinary knowledge and evidential support to negotiate ideas and engaged in argumentation more extensively during the model-based portion of the discussion.

A plausible reason for such change is that the revised model format precluded the ability to eliminate models, as the multiple claims presented in the model in spring were distinctly different from each other. Students needed to go through each of the three claims and discuss why certain claims were either more or less supported. Further, the explicit instruction on argumentation and the updated QT₅ catalyst with external, visual cues related to essential argumentation components (see Figure 2) also seemed to promote the use of certain epistemic ideals (e.g., evidentiary support) and argumentation as students evaluated each claim. For instance, at the beginning of the thin films discussion when students were evaluating the presented model, Grace used the epistemic ideal of evidentiary support and explicitly referred to the QT₅ catalyst by asking, "What have you written down as your reasoning and evidence?"

This finding is informative in terms of understanding and cultivating apt epistemic performance (i.e., "performance that achieves valuable epistemic aims through competence" [79] (p. 353)) in science classrooms. Barzilai and Chinn examined prior models of

epistemic cognition and proposed that the primary goal of epistemic education is to enable learners to achieve apt epistemic performance [79]. For the first two discussions, students used a process of elimination to approach the model-evaluation task. Using the definition of apt epistemic performance, a process of elimination could have resulted in success (i.e., choosing the correct model), but it would not have necessarily been apt (i.e., choosing the correct model by providing reasons and evidence to support all aspects of the model). Therefore, to promote students' apt epistemic performance via scientific discourse [2], it is necessary to consider the model format for discussion and encourage students to negotiate ideas about scientific phenomena using normative practices in science.

Another implication for instruction would be to optimize explicit instruction on argumentation by providing instructional tools such as a graphic organizer worksheet that visually demonstrates the components of argumentation. Such visual cues and external representation of abstract concepts can potentially support the use of argumentation components [43,66]. As students fill out these worksheets in preparation for the discussion, they are more likely to think about and negotiate ideas about these essential argumentation components during the discussion. As a result, students may be more likely to bring forth and evaluate scientific arguments by querying or providing reasoning and evidence to engage in deeper thinking about scientific phenomena.

4.3. RQ3. *The Relation of Epistemic Cognition to Authentic Questioning and Argumentation*

As evidenced across the four QT_S discussions in the current study, the relationship between the use of epistemic ideals and the initiation of authentic questions appears to be bi-directional. A plausible explanation is that when students used epistemic ideals to decide whether knowledge should be accepted as valid and found the epistemic ideal to be unmet, the dissonance prompted them to query why that knowledge claim did not meet the ideal students had in mind, and thus led to the initiation of an authentic question. On the other hand, when students responded to an authentic question, as they justified their claims by referring to their personal experience or prior knowledge as evidence, they invoked epistemic ideals accordingly.

When students invoked epistemic ideals in their responses to an authentic question, they were also likely to bring forth arguments individually or collectively, in the form of an EE, ET, or CT, indicating a close relationship between epistemic cognition and argumentation. This is possibly because the epistemic ideals students held may guide the kinds of reasoning, evidence, and arguments they brought forward and the type of disciplinary standards they used to evaluate the presented arguments. To construct an EE, students necessarily needed to build an argument by providing evidence or reasoning that met students' epistemic ideals. When students' epistemic ideals were not met, the resulting dissonance indicated a gap between what was being discussed and what students knew (see Excerpt 6) and was effective in triggering challenges during small-group discussions, which are characteristic of ET. That is, students may raise a challenge when someone says something that does not cohere with their prior knowledge. This also indicates that when students have the normative epistemic cognition knowledge necessary to critique claims, they are more likely to engage in discourse that involves thoughtful critique [30].

Finally, the relationship between epistemic ideals and CT as evidenced in Excerpt 7 reveals that even the building of knowledge without the raising of counterarguments or challenges involved the use of epistemic ideals. A possible explanation is that knowledge building involves making connections to one's prior knowledge, ideas, and explanations [80–82], as well as cognitive processes such as asking questions that probe for explanations, interpreting and evaluating information, and justifying arguments [83–85]. Such cognitive processes would require the application of epistemic ideals to help elicit or formulate responses that cohere with the students' normative disciplinary knowledge, personal experience, or other forms of prior knowledge. This also suggests the importance of teaching scientific knowledge, skills, and practices in tandem to promote student construction of knowledge through cumulative talk, as was done in our larger QT_S study [27].

4.4. Limitations

In this study, we elected to conduct a close analysis of one student group's work over the course of an instructional year to deeply examine their use of epistemic cognition and argumentation in small-group discussions. Our emphasis in this in-depth qualitative study was on ecological validity over external validity, therefore causal claims are not warranted but our findings do deeply capture epistemic cognition and argumentation in an authentic context.

Students' argumentation and epistemic cognition were observed within the context of the QT_S intervention. As part of QT_S, students received explicit instruction on argumentation and conducted regular QT_S discussions about scientific phenomena including evaluating a scientific model, and thus, these findings may not be generalized outside of this context. Instead, we argue that this study contributes to a foundation from which to further investigate students' emergent argumentation and epistemic cognition in other contexts, for example, while engaging with conflicting scientific claims [86], or to further examine how features of pedagogical practices can support students' development of argumentation practices and epistemic thinking.

5. Conclusions and Future Directions

The current study adds to the growing body of work examining situated epistemic cognition during authentic scientific practices. Within the scope of current research, we observed students' epistemic cognition through the lens of epistemic ideals and reliable processes, examined the role of contextual factors in the occurrence of epistemic cognition and argumentation, and investigated the relationship between students' epistemic cognition and their scientific argumentation during the QTs intervention. Such findings not only contribute to the field's understanding about students' epistemic cognition and argumentation in authentic science classrooms but also inform research and practice on how to develop and design effective pedagogies in ways that promote students' epistemic cognition and argumentation.

In this study, we examined students' discourse in four science discussions within one discussion group. In future work, researchers could include additional discussion groups to better capture individual and group differences (e.g., reading comprehension) [87] and explore to what extent such differences may influence students' epistemic cognition and argumentation in science. In response to the first research question, we identified various categories of epistemic ideals and reliable processes, two components of the AIR model, in students' science discourse. Future researchers could examine under what conditions students vary in their enactment of epistemic aims, and how different aims relate to scientific argumentation, to further explore students' epistemic cognition in science classrooms. Further, we found that contextual factors (e.g., model format) guided the discussion of scientific models. These contextual factors may contribute to the occurrences of different types of epistemic ideals, reliable processes, and argumentation. Researchers could extend this line of research and examine ways in which the design of a scientific modeling task can most effectively lead to students' enactment of normative scientific practices and the extent to which different attributes of the context (e.g., scientific language, background knowledge) may relate to students' scientific practices [15,88]. Finally, we identified a close relationship between students' use of epistemic ideals, authentic questioning, and argumentation across the series of QT_S discussions. This finding revealed how certain components of the QT_S intervention, such as explicit instruction on authentic questions and argumentation, may promote students' epistemic cognition. Thus, future researchers could investigate other instructional components of QT_S that might promote students' epistemic cognition in science via argumentation instruction and practice.

Indeed, in the face of various post-truth reasoning challenges, to help students develop their scientific reasoning competency and achieve "valuable epistemic aims through competence" (i.e., apt epistemic performance) [79] (p. 353), researchers likely need to work together with practitioners to design more authentic learning environments that

engage students in discussions to explore different ways of knowing and to understand how different sources of information work and why they are more or less reliable [2].

Author Contributions: Conceptualization, L.W., C.M.F., R.F.D., J.A.G. and P.K.M.; Data curation, L.W., C.M.F. and R.F.D.; Formal analysis, L.W., C.M.F., R.F.D., J.A.G. and P.K.M.; Funding acquisition, J.A.G. and P.K.M.; Investigation, L.W., C.M.F., R.F.D., J.A.G. and P.K.M.; Methodology, L.W., C.M.F., R.F.D., J.A.G. and P.K.M.; Project administration, C.M.F.; Resources, L.W., C.M.F., R.F.D., J.A.G. and P.K.M.; Supervision, J.A.G. and P.K.M.; Visualization, L.W., C.M.F. and R.F.D.; Writing—original draft, L.W., C.M.F. and R.F.D.; Writing—review and editing, L.W., C.M.F., R.F.D., J.A.G. and P.K.M. All authors have read and agreed to the published version of the manuscript.

Funding: The data reported was collected as part of a larger project funded by the National Science Foundation (USA) through Grant No. 1316347 to The Pennsylvania State University. Any opinions, findings, and conclusions or recommendations expressed are those of the author(s) and do not represent the views of the National Science Foundation.

Institutional Review Board Statement: The Pennsylvania State University’s human subjects review board approved this study, and appropriate human subjects procedures and guidelines were followed during all phases of the study and manuscript preparation.

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 participants’ informed consent in alignment with the human subjects procedures and guidelines.

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

References

- Osborne, J.; Rafanelli, S.; Kind, P. Toward a More Coherent Model for Science Education than the Crosscutting Concepts of the next Generation Science Standards: The Affordances of Styles of Reasoning. *J. Res. Sci. Teach.* **2018**, *55*, 962–981. [CrossRef]
- Chinn, C.A.; Barzilai, S.; Duncan, R.G. Education for a “Post-Truth” World: New Directions for Research and Practice. *Educ. Res.* **2021**, *50*, 51–60. [CrossRef]
- Duschl, R. Science Education in Three-Part Harmony: Balancing Conceptual, Epistemic, and Social Learning Goals. *Rev. Res. Educ.* **2008**, *32*, 268–291. [CrossRef]
- National Research Council. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*; National Academies Press: Cambridge, MA, USA, 2012. [CrossRef]
- Greene, J.A.; Sandoval, W.A.; Bråten, I. *Handbook of Epistemic Cognition*; Routledge: Abingdon, UK, 2016. [CrossRef]
- Greene, J.A.; Cartiff, B.M.; Duke, R.F. A Meta-Analytic Review of the Relationship between Epistemic Cognition and Academic Achievement. *J. Educ. Psychol.* **2018**, *110*, 1084–1111. [CrossRef]
- Tang, K.-S. The Use of Epistemic Tools to Facilitate Epistemic Cognition & Metacognition in Developing Scientific Explanation. *Cogn. Instr.* **2020**, *38*, 474–502. [CrossRef]
- Ke, L.; Schwarz, C.V. Using Epistemic Considerations in Teaching: Fostering Students’ Meaningful Engagement in Scientific Modeling. In *Towards a Competence-Based View on Models and Modeling in Science Education*; Upmeyer zu Belzen, A., Krüger, D., van Driel, J., Eds.; Models and Modeling in Science Education; Springer International Publishing: Cham, Switzerland, 2019; Volume 12, pp. 181–199. [CrossRef]
- Sinatra, G.M.; Chinn, C.A. Thinking and Reasoning in Science: Promoting Epistemic Conceptual Change. In *APA Educational Psychology Handbook, Vol 3: Application to Learning and Teaching*; Harris, K.R., Graham, S., Urdan, T., Bus, A.G., Major, S., Swanson, H.L., Eds.; American Psychological Association: Washington, DC, USA, 2012; pp. 257–282. [CrossRef]
- Sandoval, W. Disciplinary Insights into the Study of Epistemic Cognition. In *Handbook of epistemic cognition*; Greene, J.A., Sandoval, W.A., Bråten, I., Eds.; Routledge: Abingdon, UK, 2016; pp. 184–194.
- Kienhues, D.; Jucks, R.; Bromme, R. Sealing the Gateways for Post-Truthism: Reestablishing the Epistemic Authority of Science. *Educ. Psychol.* **2020**, *55*, 144–154. [CrossRef]
- Reith, M.; Nehring, A. Scientific Reasoning and Views on the Nature of Scientific Inquiry: Testing a New Framework to Understand and Model Epistemic Cognition in Science. *Int. J. Sci. Educ.* **2020**, *42*, 2716–2741. [CrossRef]
- Cartiff, B.M.; Duke, R.F.; Greene, J.A. The Effect of Epistemic Cognition Interventions on Academic Achievement: A Meta-Analysis. *J. Educ. Psychol.* **2021**, *113*, 477–498. [CrossRef]
- Osborne, J. The 21st Century Challenge for Science Education: Assessing Scientific Reasoning. *Think. Skills Creat.* **2013**, *10*, 265–279. [CrossRef]
- Chinn, C.; Sandoval, W. Epistemic Cognition and Epistemic Development. In *International Handbook of the Learning Sciences*; Fischer, F., Hmelo-Silver, C.E., Goldman, S.R., Reimann, P., Eds.; Routledge: New York, NY, USA, 2018; pp. 24–33. [CrossRef]

16. Beniermann, A.; Mecklenburg, L.; Upmeier zu Belzen, A. Reasoning on Controversial Science Issues in Science Education and Science Communication. *Educ. Sci.* **2021**, *11*, 522. [CrossRef]
17. Upmeier zu Belzen, A.; Engelschalt, P.; Krüger, D. Modeling as Scientific Reasoning—The Role of Abductive Reasoning for Modeling Competence. *Educ. Sci.* **2021**, *11*, 495. [CrossRef]
18. Mason, L.; Scirica, F. Prediction of Students' Argumentation Skills about Controversial Topics by Epistemological Understanding. *Learn. Instr.* **2006**, *16*, 492–509. [CrossRef]
19. Nussbaum, E.M.; Sinatra, G.M.; Poliquin, A. Role of Epistemic Beliefs and Scientific Argumentation in Science Learning. *Int. J. Sci. Educ.* **2008**, *30*, 1977–1999. [CrossRef]
20. Weinstock, M.P. Cognitive Bases for Effective Participation in Democratic Institutions: Argument Skill and Juror Reasoning. *Theory Res. Soc. Educ.* **2005**, *33*, 73–102. [CrossRef]
21. Weinstock, M. Knowledge-Telling and Knowledge-Transforming Arguments in Mock Jurors' Verdict Justifications. *Think. Reason.* **2011**, *17*, 282–314. [CrossRef]
22. Weinstock, M.; Cronin, M.A. The Everyday Production of Knowledge: Individual Differences in Epistemological Understanding and Juror-Reasoning Skill. *Appl. Cognit. Psychol.* **2003**, *17*, 161–181. [CrossRef]
23. Duncan, R.G.; Chinn, C.A. New Directions for Research on Argumentation: Insights from the AIR Framework for Epistemic Cognition. *Z. Pädagog. Psychol.* **2016**, *30*, 155–161. [CrossRef]
24. Herrenkohl, L.R.; Cornelius, L. Investigating Elementary Students' Scientific and Historical Argumentation. *J. Learn. Sci.* **2013**, *22*, 413–461. [CrossRef]
25. Iordanou, K.; Constantinou, C.P. Supporting Use of Evidence in Argumentation Through Practice in Argumentation and Reflection in the Context of SOCRATES Learning Environment: Supporting use of evidence in argumentation. *Sci. Educ.* **2015**, *99*, 282–311. [CrossRef]
26. Murphy, P.K.; Wilkinson, I.A.G.; Soter, A.O.; Hennessey, M.N.; Alexander, J.F. Examining the Effects of Classroom Discussion on Students' Comprehension of Text: A Meta-Analysis. *J. Educ. Psychol.* **2009**, *101*, 740–764. [CrossRef]
27. Murphy, P.K.; Greene, J.A.; Allen, E.; Baszczewski, S.; Swearingen, A.; Wei, L.; Butler, A.M. Fostering High School Students' Conceptual Understanding and Argumentation Performance in Science through *Quality Talk* Discussions. *Sci. Educ.* **2018**, *102*, 1239–1264. [CrossRef]
28. Reznitskaya, A.; Gregory, M. Student Thought and Classroom Language: Examining the Mechanisms of Change in Dialogic Teaching. *Educ. Psychol.* **2013**, *48*, 114–133. [CrossRef]
29. Chinn, C.A.; Buckland, L.A.; Samarapungavan, A. Expanding the Dimensions of Epistemic Cognition: Arguments from Philosophy and Psychology. *Educ. Psychol.* **2011**, *46*, 141–167. [CrossRef]
30. Chinn, C.A.; Rinehart, R.W.; Buckland, L.A. Epistemic Cognition and Evaluating Information. In *Processing Inaccurate Information*; Rapp, D.N., Braasch, J.L.G., Eds.; The MIT Press: Cambridge, MA, USA, 2014; pp. 425–453. [CrossRef]
31. Muis, K.R. The Role of Epistemic Beliefs in Self-Regulated Learning. *Educ. Psychol.* **2007**, *42*, 173–190. [CrossRef]
32. Rosenberg, S.; Hammer, D.; Phelan, J. Multiple Epistemological Coherences in an Eighth-Grade Discussion of the Rock Cycle. *J. Learn. Sci.* **2006**, *15*, 261–292. [CrossRef]
33. Barzilai, S. "Half-Reliable": A Qualitative Analysis of Epistemic Thinking in and about a Digital Game. *Contemp. Educ. Psychol.* **2017**, *51*, 51–66. [CrossRef]
34. Bricker, L.A.; Bell, P. Conceptualizations of Argumentation from Science Studies and the Learning Sciences and Their Implications for the Practices of Science Education. *Sci. Educ.* **2008**, *92*, 473–498. [CrossRef]
35. Duschl, R.A.; Osborne, J. Supporting and Promoting Argumentation Discourse in Science Education. *Stud. Sci. Educ.* **2002**, *38*, 39–72. [CrossRef]
36. Kuhn, D. Science as Argument: Implications for Teaching and Learning Scientific Thinking. *Sci. Educ.* **1993**, *77*, 319–337. [CrossRef]
37. van Eemeren, F.H.; Grootendorst, R. Fundamentals of Argumentation Theory: A Handbook of Historical Backgrounds and Contemporary Developments. *College Compos. Commun.* **1997**, *48*, 437–438. [CrossRef]
38. Driver, R.; Osborne, J.; Newton, P. Establishing the Norms of Scientific Argumentation in Classrooms. *Sci. Educ.* **2000**, *84*, 287–312. [CrossRef]
39. Walton, D. *Informal Logic: A Pragmatic Approach*; Cambridge University Press: Cambridge, UK, 2008.
40. Osborne, J. Arguing to Learn in Science: The Role of Collaborative, Critical Discourse. *Science* **2010**, *328*, 463–466. [CrossRef] [PubMed]
41. Asterhan, C.S.C.; Schwarz, B.B. Argumentation for Learning: Well-Trodden Paths and Unexplored Territories. *Educ. Psychol.* **2016**, *51*, 164–187. [CrossRef]
42. Chinn, C.A.; Clark, D.B. Learning Through Collaborative Argumentation. In *The International Handbook of Collaborative Learning*; Hmelo-Silver, C.E., Chinn, C., Chan, C.K.K., O'Donnell, A., Eds.; Routledge: Abingdon, UK, 2011; pp. 307–324. [CrossRef]
43. Nussbaum, E.M. Collaborative Discourse, Argumentation, and Learning: Preface and Literature Review. *Contemp. Educ. Psychol.* **2008**, *3*, 345–359. [CrossRef]
44. Jiménez-Aleixandre, M.P.; Brocos, P. Shifts in Epistemic Status in Argumentation and in Conceptual Change. In *Converging Perspectives on Conceptual Change*; Amin, T.G., Levrini, O., Eds.; Routledge: Abingdon, UK, 2017; pp. 171–179. [CrossRef]

45. Hofer, B.K.; Bendixen, L.D. Personal Epistemology: Theory, Research, and Future Directions. In *APA Educational Psychology Handbook, Vol 1: Theories, Constructs, and Critical Issues*; Harris, K.R., Graham, S., Urdan, T., McCormick, C.B., Sinatra, G.M., Sweller, J., Eds.; American Psychological Association: Washington, DC, USA, 2012; pp. 227–256. [CrossRef]
46. Iordanou, K.; Kendeou, P.; Beker, K. Argumentative Reasoning. In *Handbook of Epistemic Cognition*; Greene, J.A., Sandoval, W.A., Bråten, I., Eds.; Routledge: Abingdon, UK, 2016; pp. 39–53.
47. Mason, L.; Boscolo, P. Role of Epistemological Understanding and Interest in Interpreting a Controversy and in Topic-Specific Belief Change. *Contemp. Educ. Psychol.* **2004**, *29*, 103–128. [CrossRef]
48. Weinstock, M.P.; Neuman, Y.; Glassner, A. Identification of Informal Reasoning Fallacies as a Function of Epistemological Level, Grade Level, and Cognitive Ability. *J. Educ. Psychol.* **2006**, *98*, 327–341. [CrossRef]
49. Nussbaum, E.M.; Bendixen, L.D. Approaching and Avoiding Arguments: The Role of Epistemological Beliefs, Need for Cognition, and Extraverted Personality Traits. *Contemp. Educ. Psychol.* **2003**, *28*, 573–595. [CrossRef]
50. Kuhn, D.; Zillmer, N.; Crowell, A.; Zavala, J. Developing Norms of Argumentation: Metacognitive, Epistemological, and Social Dimensions of Developing Argumentative Competence. *Cogn. Instr.* **2013**, *31*, 456–496. [CrossRef]
51. Muis, K.R.; Trevors, G.; Chevrier, M. Epistemic Climate for Epistemic Change. In *Handbook of Epistemic Cognition*; Greene, J.A., Sandoval, W.A., Bråten, I., Eds.; Routledge: Abingdon, UK, 2016; pp. 331–359.
52. Dyer, M.K.; Moynihan, C. *Open-Ended Question in Elementary Mathematics: Instruction & Assessment*; Eye on Education: Rockville, MD, USA, 2000.
53. Hancock, C.L. Implementing the Assessment Standards for School Mathematics: Enhancing Mathematics Learning with Open-Ended Questions. *MT* **1995**, *88*, 496–499. [CrossRef]
54. Buchheister, K.; Jackson, C.; Taylor, C.E. What, How, Who: Developing Mathematical Discourse. *MTMS* **2019**, *24*, 202–208. [CrossRef]
55. Wilkinson, I.A.G.; Soter, A.O.; Murphy, P.K. Developing a Model of Quality Talk about Literary Text. In *Bringing Reading Research to Life*; McKeown, M.G., Kucan, L., Eds.; Guilford Press: New York, NY, USA, 2009; pp. 142–169.
56. Murphy, P.K.; Rowe, M.L.; Ramani, G.; Silverman, R. Promoting Critical-Analytic Thinking in Children and Adolescents at Home and in School. *Educ. Psychol. Rev.* **2014**, *26*, 561–578. [CrossRef]
57. Wei, L.; Murphy, P.K.; Firetto, C.M. How Can Teachers Facilitate Productive Small-Group Talk? An Integrated Taxonomy of Teacher Discourse Moves. *Elem. School J.* **2018**, *118*, 578–609. [CrossRef]
58. Murphy, P.K.; Firetto, C.M. Quality Talk: A Blueprint for Productive Talk. In *Classroom Discussions in Education*; Murphy, P.K., Ed.; Routledge: Abingdon, UK, 2017; pp. 101–133.
59. Murphy, P.K.; Firetto, C.M.; Wei, L.; Li, M.; Croninger, R.M.V. What REALLY Works: Optimizing Classroom Discussions to Promote Comprehension and Critical-Analytic Thinking. *Policy Insights Behav. Brain Sci.* **2016**, *3*, 27–35. [CrossRef]
60. Murphy, P.K.; Greene, J.A.; Firetto, C.M.; Hendrick, B.D.; Li, M.; Montalbano, C.; Wei, L. Quality Talk: Developing Students' Discourse to Promote High-Level Comprehension. *Am. Educ. Res. J.* **2018**, *55*, 1113–1160. [CrossRef]
61. Lloyd, G.M.; Murphy, P.K. Mathematical Argumentation in Small-Group Discussions of Complex Mathematical Tasks in Elementary Teacher Education Settings (accepted). In *Mathematical Challenges for All*; Leikin, R., Ed.; Springer: Berlin/Heidelberg, Germany.
62. Wei, L.; Murphy, P.K.; Wu, S. Recontextualizing Quality Talk for an Eighth-Grade English Classroom in China. *ECNU Rev. Educ.* **2020**. [CrossRef]
63. Murphy, P.K.; The Quality Talk Team. From Theoretical Roots to Empirical Outcomes: The Interdisciplinary Foundations of Quality Talk in Taiwan. In *The Theory and Practice of Group Discussion with Quality Talk*; Learning Sciences for Higher Education; Chen, C.-C., Lo, M.-L., Eds.; Springer: Singapore, 2021; pp. 1–21. [CrossRef]
64. Wei, L.; Murphy, P.K. Recontextualising Discourse-Intensive Interventions for Multilingual Contexts: Implementing Quality Talk in China. In *Multilingualism in the Classroom*; Omidire, F., Ed.; UCT Press: Cape Town, South Africa, 2019; pp. 57–81.
65. Murphy, P.K.; Ebersöhn, L.; Omidire, F.; Firetto, C.M. Exploring the Structure and Content of Discourse in Remote, Rural South African Classrooms. *SAJE* **2020**, *40*, S1–S11. [CrossRef]
66. Wei, L.; Firetto, C.M.; Murphy, P.K.; Li, M.; Greene, J.A.; Croninger, R.M.V. Facilitating Fourth-Grade Students' Written Argumentation: The Use of an Argumentation Graphic Organizer. *J. Educ. Res.* **2019**, *112*, 627–639. [CrossRef]
67. Schwarz, C.V.; Reiser, B.J.; Davis, E.A.; Kenyon, L.; Acher, A.; Fortus, D.; Shwartz, Y.; Hug, B.; Krajcik, J. Developing a Learning Progression for Scientific Modeling: Making Scientific Modeling Accessible and Meaningful for Learners. *J. Res. Sci. Teach.* **2009**, *46*, 632–654. [CrossRef]
68. Schwarz, C.; Reiser, B.J.; Acher, A.; Kenyon, L.; Fortus, D. MoDeLS. In *Learning Progressions in Science*; Alonzo, A.C., Gotwals, A.W., Alonzo, A.C., Gotwals, A., Eds.; SensePublishers: Rotterdam, The Netherlands, 2012; pp. 101–137. [CrossRef]
69. Schwarz, C.V.; White, B.Y. Metamodeling Knowledge: Developing Students' Understanding of Scientific Modeling. *Cogn. Instr.* **2005**, *23*, 165–205. [CrossRef]
70. Pluta, W.J.; Chinn, C.A.; Duncan, R.G. Learners' Epistemic Criteria for Good Scientific Models. *J. Res. Sci. Teach.* **2011**, *48*, 486–511. [CrossRef]
71. Marshall, C.; Rossman, G.B. *Designing Qualitative Research*, 6th ed.; SAGE Publications, Inc.: London, UK, 2015.
72. Murphy, P.K.; Firetto, C.M.; Greene, J.A.; Butler, A.M. *Analyzing the Talk in Quality Talk Discussions: A Coding Manual*; The Pennsylvania State University: State College, PA, USA, 2017. [CrossRef]

73. van Dijk, T.A. Episodes as Units of Discourse Analysis. In *Analyzing Discourse: Text and Talk*; Tannen, D., Ed.; Georgetown University Press: Washington, DC, USA, 1981; pp. 177–195.
74. Michaels, S.; O'Connor, C.; Resnick, L.B. Deliberative Discourse Idealized and Realized: Accountable Talk in the Classroom and in Civic Life. *Stud. Philos. Educ.* **2008**, *27*, 283–297. [CrossRef]
75. Hogan, K.; Maglienti, M. Comparing the Epistemological Underpinnings of Students' and Scientists' Reasoning about Conclusions. *J. Res. Sci. Teach.* **2001**, *38*, 663–687. [CrossRef]
76. Samarapungavan, A. Children's Judgments in Theory Choice Tasks: Scientific Rationality in Childhood. *Cognition* **1992**, *45*, 1–32. [CrossRef]
77. Moshman, D.; Tarricone, P. Logical and Causal Reasoning. In *Handbook of Epistemic Cognition*; Greene, J.A., Sandoval, W.A., Bråten, I., Eds.; Routledge: Abingdon, UK, 2016; pp. 54–67.
78. Kienhues, D.; Ferguson, L.; Stahl, E. Diverging Information and Epistemic Change. In *Handbook of Epistemic Cognition*; Greene, J.A., Sandoval, W.A., Bråten, I., Eds.; Routledge: Abingdon, UK, 2016; pp. 318–330.
79. Barzilai, S.; Chinn, C.A. On the Goals of Epistemic Education: Promoting Apt Epistemic Performance. *J. Learn. Sci.* **2018**, *27*, 353–389. [CrossRef]
80. Kolodner, J.L.; Camp, P.J.; Crismond, D.; Fasse, B.; Gray, J.; Holbrook, J.; Puntambekar, S.; Ryan, M. Problem-Based Learning Meets Case-Based Reasoning in the Middle-School Science Classroom: Putting Learning by Design(TM) Into Practice. *J. Learn. Sci.* **2003**, *12*, 495–547. [CrossRef]
81. Krajcik, J.; McNeill, K.L.; Reiser, B.J. Learning-Goals-Driven Design Model: Developing Curriculum Materials That Align with National Standards and Incorporate Project-Based Pedagogy: Learning-Goals-Driven Design. *Sci. Educ.* **2008**, *92*, 1–32. [CrossRef]
82. von Glasersfeld, E. *Radical Constructivism a Way of Knowing and Learning*; Falmer: Brighton, UK, 1995.
83. Chan, C.K.K.; Burtis, P.J.; Scardamalia, M.; Bereiter, C. Constructive Activity in Learning from Text. *Am. Educ. Res. J.* **1992**, *29*, 97–118. [CrossRef]
84. King, A.; Rosenshine, B. Effects of Guided Cooperative Questioning on Children's Knowledge Construction. *J. Exp. Educ.* **1993**, *61*, 127–148. [CrossRef]
85. van Aalst, J. Distinguishing Knowledge-Sharing, Knowledge-Construction, and Knowledge-Creation Discourses. *Computer Suppor. Learn.* **2009**, *4*, 259–287. [CrossRef]
86. Tabak, I. Functional Scientific Literacy: Disciplinary Literacy Meets Multiple Source Use. In *Handbook of Multiple Source Use*; Braasch, J.L.G., Bråten, I., McCrudden, M.T., Eds.; Routledge: Abingdon, UK, 2018; pp. 221–237.
87. Schlatter, E.; Lazonder, A.W.; Molenaar, I.; Janssen, N. Individual Differences in Children's Scientific Reasoning. *Educ. Sci.* **2021**, *11*, 471. [CrossRef]
88. Gilbert, J.K. On the Nature of "Context" in Chemical Education. *Int. J. Sci. Educ.* **2006**, *28*, 957–976. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Education Sciences Editorial Office
E-mail: education@mdpi.com
www.mdpi.com/journal/education



MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34

www.mdpi.com



ISBN 978-3-0365-4547-9