THE EMERGENCE AND DEVELOPMENT OF SCIENTIFIC THINKING DURING THE EARLY YEARS: BASIC PROCESSES AND SUPPORTIVE CONTEXTS

EDITED BY: Ageliki Nicolopoulou, Amanda C. Brandone, Stella Vosniadou and Christopher Osterhaus

PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-88966-642-3
DOI 10.3389/978-2-88966-642-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

THE EMERGENCE AND DEVELOPMENT OF SCIENTIFIC THINKING DURING THE EARLY YEARS: BASIC PROCESSES AND SUPPORTIVE CONTEXTS

Topic Editors:
Ageliki Nicolopoulou, Lehigh University, United States
Amanda C. Brandone, Lehigh University, United States
Stella Vosniadou, Flinders University, Australia
Christopher Osterhaus, University of Vechta, Germany

Citation: Nicolopoulou, A., Brandone, A. C., Vosniadou, S., Osterhaus, C., eds. (2021). The Emergence and Development of Scientific Thinking during the Early Years: Basic Processes and Supportive Contexts. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-642-3

Table of Contents

05 ***Editorial: The Emergence and Development of Scientific Thinking During the Early Years: Basic Processes and Supportive Contexts***
Christopher Osterhaus, Amanda C. Brandone, Stella Vosniadou and Ageliki Nicolopoulou

10 ***Individual Differences in Children's Development of Scientific Reasoning Through Inquiry-Based Instruction: Who Needs Additional Guidance?***
Erika Schlatter, Inge Molenaar and Ard W. Lazonder

24 ***Scientific Reasoning in Biology – the Impact of Domain-General and Domain-Specific Concepts on Children's Observation Competency***
Janina Klemm, Pamela Flores, Beate Sodian and Birgit J. Neuhaus

36 ***Children's Spontaneous Gestures Reflect Verbal Understanding of the Day/Night Cycle***
Caroline M. Gaudreau, Florencia K. Anggoro and Benjamin D. Jee

43 ***Parental Pre-knowledge Enhances Guidance During Inquiry-Based Family Learning in a Museum Context: An Individual Differences Perspective***
Rooske K. Franse, Tessa J. P. Van Schijndel and Maartje E. J. Raijmakers

59 ***Embedding Scientific Explanations Into Storybooks Impacts Children's Scientific Discourse and Learning***
Kathryn A. Leech, Amanda S. Haber, Youmna Jalkh and Kathleen H. Corriveau

71 ***Theory of Mind, Personal Epistemology, and Science Learning: Exploring Common Conceptual Components***
Natassa Kyriakopoulou and Stella Vosniadou

86 ***How Children's Cognitive Reflection Shapes Their Science Understanding***
Andrew G. Young and Andrew Shtulman

94 ***Young Schoolchildren's Epistemic Development: A Longitudinal Qualitative Study***
Michael Weinstock, Vardit Israel, Hadas Fisher Cohen, Iris Tabak and Yifat Harari

107 ***The Impact of a Construction Play on 5- to 6-Year-Old Children's Reasoning About Stability***
Anke Maria Weber, Timo Reuter and Miriam Leuchter

127 ***Nascent Inquiry, Metacognitive, and Self-Regulation Capabilities Among Preschoolers During Scientific Exploration***
Ronit Fridman, Sigal Eden and Ornit Spektor-Levy

143 ***Hands-On Exploration of Cubes' Floating and Sinking Benefits Children's Subsequent Buoyancy Predictions***
Johanna E. van Schaik, Tessa Slim, Rooske K. Franse and Maartje E. J. Raijmakers

161 ***Inquiry-Based Learning and Conceptual Change in Balance Beam Understanding***
Joep van der Graaf

177 ***Preschoolers' Induction of the Concept of Material Kind to Make Predictions: The Effects of Comparison and Linguistic Labels***
Ilonca Hardy, Henrik Saalbach, Miriam Leuchter and Lennart Schalk

191 ***Investigating Science Together: Inquiry-Based Training Promotes Scientific Conversations in Parent-Child Interactions***
Ian L. Chandler-Campbell, Kathryn A. Leech and Kathleen H. Corriveau

203 ***Supporting Early Scientific Thinking Through Curiosity***
Jamie J. Jirout

210 ***Family Conversations About Heat and Temperature: Implications for Children's Learning***
Megan R. Luce and Maureen A. Callanan

228 ***Of Blickets, Butterflies, and Baby Dinosaurs: Children's Diagnostic Reasoning Across Domains***
Deena Skolnick Weisberg, Elysia Choi and David M. Sobel

241 ***Development of Conceptual Flexibility in Intuitive Biology: Effects of Environment and Experience***
Nicole Betz and John D. Coley

260 ***Reasoning With Conditionals About Everyday and Mathematical Concepts in Primary School***
Anastasia Datsogianni, Beate Sodian, Henry Markovits and Stefan Ufer

276 ***Prompting Children's Belief Revision About Balance Through Primary and Secondary Sources of Evidence***
Nicole E. Larsen, Vaunam P. Venkadasalam and Patricia A. Ganea

288 ***The Predictive Value of Children's Understanding of Indeterminacy and Confounding for Later Mastery of the Control-of-Variables Strategy***
Sonja Peteranderl and Peter A. Edelsbrunner

frontiers in Psychology

EDITORIAL
published: 19 February 2021
doi: 10.3389/fpsyg.2021.629384

Editorial: The Emergence and Development of Scientific Thinking During the Early Years: Basic Processes and Supportive Contexts

Christopher Osterhaus[1†], *Amanda C. Brandone*[2], *Stella Vosniadou*[3] and *Ageliki Nicolopoulou*[2*]

[1] Department of Psychology, Ludwig Maximillian University of Munich, Munich, Germany, [2] Department of Psychology, Lehigh University, Bethlehem, PA, United States, [3] College of Education, Psychology and Social Work, Flinders University, Adelaide, SA, Australia

Keywords: scientific reasoning, science concepts, science education, parent-child interaction, domain-general, domain-specific, inquiry-based learning, conceptual change

OPEN ACCESS

Edited and reviewed by:
Katharina J. Rohlfing, University of Paderborn, Germany

***Correspondence:**
Ageliki Nicolopoulou
agn3@lehigh.edu

†Present address:
Christopher Osterhaus, University of Vechta, Faculty of Education and Social Sciences, Vechta, Germany

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *14 November 2020*
Accepted: *28 January 2021*
Published: *19 February 2021*

Citation:
Osterhaus C, Brandone AC, Vosniadou S and Nicolopoulou A (2021) Editorial: The Emergence and Development of Scientific Thinking During the Early Years: Basic Processes and Supportive Contexts. Front. Psychol. 12:629384. doi: 10.3389/fpsyg.2021.629384

Editorial on the Research Topic

The Emergence and Development of Scientific Thinking During the Early Years: Basic Processes and Supportive Contexts

Scientific ways of knowing and the ability to think and act to enhance our understanding of the natural and social world are among the greatest human intellectual achievements. In modern knowledge-rich societies, scientific thinking is of crucial importance since it allows participation in increasingly complex public discourse and informed decision-making about socio-scientific issues (Sadler, 2004), such as climate change or health crises. Accordingly, many regard scientific thinking to be a critical 21st century skill which should be fostered from early in development (Trilling and Fadel, 2009).

Early developmental and educational researchers have been skeptical about young children's science competencies—including both early *science understanding* (i.e., the knowledge about scientific explanations of phenomena across various domains) and early *scientific reasoning* (i.e., the reasoning process involved in the construction of science knowledge), which together make up scientific thinking (Kuhn, 2002). In particular, research on science understanding has shown how difficult it is for children to develop scientific explanations of the world when they are in conflict with children's intuitive understandings (Vosniadou and Brewer, 1992); research on scientific reasoning has shown that young children tend to be unsystematic in the experimentation strategies they use (Tschirgi, 1980), forget to keep track of inquiry goals, experiments, and outcomes (Kuhn et al., 2008), and try to produce rather than to investigate causal effects (Lehrer and Schauble, 2007). Although difficulties persist throughout elementary school (e.g., Kuhn et al., 1995) and even among adults (Wason, 1968), recent research has shown that a broad range of scientific thinking skills are present earlier than previously expected (Sodian et al., 1991; Ruffman et al., 1993; van der Graaf et al., 2015; Köksal-Tuncer and Sodian, 2018; Koerber and Osterhaus, 2019).

The present Research Topic adds to these findings by bringing together cutting-edge research on both the basic abilities in early childhood that form the foundation of mature scientific thinking as well as the contexts, strategies, and processes that support its development. Specifically, articles in this Research Topic explore (1) the family interactions at home and in museum contexts that encourage the development of early science concepts, (2) the domain-specific and domain-general cognitive processes involved in the acquisition of scientific thinking, (3) the development and facilitation of scientific reasoning, and (4) how and under which conditions scientific thinking can

promote the acquisition and revision of science knowledge.

FAMILY INTERACTIONS AT HOME AND IN MUSEUM CONTEXTS AROUND SCIENCE CONCEPTS

Informal everyday interactions between parents and children are the earliest contexts that help children develop their science knowledge, taking advantage of their curiosity about the physical world, which Jirout considers critical for promoting science knowledge in young children. In recent years, there has been increased interest in these informal everyday contexts, which is also reflected in the articles in this Research Topic.

Luce and Callanan analyzed parents' everyday conversations about heat and temperature with 2–6-year-olds drawn from the CHILDES language database and from a parent-child book-reading study. This study highlights the need for detailed investigations of everyday verbal input that children receive to shed further light on how children build intuitive concepts about science phenomena.

The rest of the articles in this section focus on how to best structure parental interactions to encourage the development of science understanding, especially at informal museum science exhibits. Leech et al. provide evidence that having parents read children's books that include mechanistic explanations about science concepts to their 4- and 5-year-olds prompted these dyads to use more mechanistic language and to more successfully solve a related science problem. Similarly, Franse et al. found that when parents had pre-knowledge of a task presented in a museum exhibit by being shown the solution, they were better able to scaffold their 8–12-year-olds than parents who did not have pre-knowledge: the parents with pre-knowledge interacted longer with their children, asked more open-ended questions, led their children to inquire on their own, and were less likely to interpret the results for them than parents with no pre-knowledge. Finally, Chandler-Campbell et al. demonstrated that parents who participated in an *inquiry-based* rather than a *statement-sharing* intervention were able to better leverage children's curiosity: these parents asked more questions, including causal ones, and children provided more scientific content in response. Taken together, these studies indicate that even in informal museum contexts the best results in promoting science knowledge require parental guidance and inquiry-based scaffolding from knowledgeable parents.

THE ROLE OF DOMAIN-SPECIFIC AND DOMAIN-GENERAL COGNITIVE PROCESSES

While in the earliest contexts, parents play an important role in promoting children's science learning, with age children become increasingly able to independently engage in scientific reasoning. Scientific reasoning is defined as intentional knowledge seeking (Kuhn, 2002). As such, developing basic scientific reasoning skills requires, first and foremost, that children recognize that knowledge is not simply a copy of the external world, but rather that knowledge claims are judgments evaluated in light of evidence (Kuhn et al., 2008). Once children attain this foundational insight, they develop a more complex understanding of how the human mind works and, in turn, are ready to acquire a more advanced understanding of the nature of science (Osterhaus et al., 2017), including the realization that scientists feel certain about their beliefs when their beliefs are supported by their interpretations of evidence.

Children's understanding of the constructive nature of knowledge has been studied extensively by researchers interested in (advanced) theory of mind and is also examined in relation to scientific reasoning in this Research Topic. In a 2-year longitudinal study involving 7–8-year-old children, Weinstock et al. found a link between children's early understanding of the interpretative nature of the mind and their epistemological understanding. That is, children who better understood that representations of the external world result from our mind's active interpretation, were more likely to apply that knowledge in a scientific context (epistemological understanding). Kyriakopoulou and Vosniadou showed that higher-order false belief reasoning (i.e., recognizing that someone may hold a false belief about a belief) in 10–12-year-olds was associated with more mature epistemological beliefs and more advanced knowledge about observational astronomy.

In addition to theory of mind, ample metacognition and better information-processing skills are also necessary to develop advanced scientific thinking. Consistent with this emphasis on domain-general cognitive processes, Betz and Coley demonstrated that conceptual flexibility (i.e., the ability to switch between different ways of organizing knowledge) increased with both age and experience, with implications for children's biological knowledge. Young and Shtulman reported that 5–12-year-olds' cognitive reflection (defined as the tendency to reflect on one's own thinking) was related to children's understanding of counterintuitive science ideas. And Fridman et al. demonstrated the impact of metacognition and self-regulation on 5–6-year-olds' scientific exploration. Taken together, these studies highlight the ways in which cognitive abilities outside the realm of science (e.g., metacognition, self-regulation, conceptual flexibility, theory of mind) are associated with children's ability to reason scientifically.

DEVELOPMENT AND FACILITATION OF DIVERSE FORMS OF SCIENTIFIC REASONING

Children's exploration of science phenomena has primarily been investigated by asking how proficient they are in experimentation using the control-of-variables strategy (CVS). CVS holds that informative experiments must only vary a single variable at a time while keeping all others constant, and as such, allows for systematic exploration of cause-effect relations. A consistent finding in the literature is that CVS is difficult for elementary school children (e.g., Kuhn et al., 1995; Bullock and Ziegler, 1999; Croker and Buchanan, 2011). However, experimentation is not the only mode of scientific reasoning (Kind and Osborne,

2017), and studies in this Research Topic highlight several other important forms of scientific reasoning, including some that develop much earlier than the ability to control variables.

Klemm et al. found that 4–6-year-olds have competencies in observation that go beyond merely making observations and include diverse epistemic activities, such as asking questions about, testing, or making sense of observations. Weisberg et al. demonstrated that 4–10-year-olds are capable of diagnostic reasoning (i.e., the ability to infer causes from systematic observations of patterns of data about cause-effect relations) across multiple contexts. Datsogianni et al. reported abilities in conditional reasoning (i.e., reasoning about if-then statements) in both familiar and mathematical contexts in children aged 7–12 years. And Peteranderl and Edelsbrunner demonstrated important precursors to the development of CVS in 9–11-year-olds, including an understanding of indeterminacy (whether available evidence is sufficient to warrant a conclusion) and confounding (whether confounding variables are appropriately controlled). In sum, although observation, data interpretation, conditional or diagnostic reasoning, and experimentation may vary in complexity and age of acquisition, the studies in this section highlight how each of these processes can be considered genuine forms of scientific reasoning.

Importantly, Schlatter et al. showed that these abilities are also responsive to intervention. In particular, a 5-week training program resulted in significant competence gains across diverse scientific reasoning abilities (i.e., hypothesis formation, experimentation, and data interpretation). Whether or not these effects are long-lasting, and whether individual differences in children's performance on different scientific reasoning measures are stable over time, remain open questions.

SCIENTIFIC THINKING IN THE SERVICE OF PROMOTING CHANGES IN SCIENCE KNOWLEDGE

The scientific reasoning processes described above can be applied in the service of the acquisition of science knowledge. A large body of research has shown that children have intuitive knowledge in various science domains, including physics, biology, and astronomy (Shtulman and Walker, 2020), and that they bring this intuitive knowledge to the task of science learning (Vosniadou, 2019). Thus, the development of accurate science knowledge often involves a gradual process of assimilating new knowledge and revising or replacing prior scientifically-incorrect beliefs (Vosniadou, 2002). Several papers in this Research Topic examine these processes, shedding light on the conditions under which the revision of science knowledge takes place and how it can be promoted and assessed.

Van Schaik et al. found that providing systematic evidence highlighting key variables (vs. non-systematic evidence) promoted 4–9-year-olds' predictions and explanations about buoyancy events. Hardy et al. showed that conditions designed to facilitate comparison also enhanced 4–7-year-olds' predictions about buoyancy. Larsen et al. and Weber et al. explored how the nature of the evidence children receive impacts revisions to their science knowledge. Larsen et al. found that 5-year-olds' learning about balance was facilitated following anomalous evidence experienced directly (through a hand-on task) and indirectly (through illustrations in a picture book) compared to a control condition. Moreover, Weber et al. found that children who received a combination of both verbal and material forms of scaffolding in a play-based intervention were most likely to adjust their theories about balance in the face of counterevidence. Van der Graaf further demonstrated how an inquiry-based lesson involving generating hypotheses and gathering and evaluating evidence promoted conceptual change as revealed by changes in 8–13-year-olds' strategy use on a balance beam task. Finally, Gaudreau et al. found that the gestures third-graders use when describing the day/night cycle can provide insight into their developing science knowledge by reflecting children's current understandings and potentially foreshadowing future conceptual change.

Findings from across these studies reveal advances and revisions in children's science knowledge that are driven by salient and often hands-on learning experiences. Children can engage in belief revision when they are confronted with systematic evidence (including evidence that conflicts with their prior beliefs) and when they are encouraged to reflect on or explain that evidence. Consistent with the above findings about parent-child interactions and the broader literature on inquiry-based learning (e.g., Lazonder and Harmsen, 2016), the studies in this section also suggest that learning and belief revision are more successful when the task includes greater structure and specific guidance. Finally, these studies highlight the various ways in which children's initial knowledge can facilitate or constrain their ability to learn from interventions.

CONCLUDING REMARKS AND FUTURE DIRECTIONS

The wide range of studies in this Research Topic from both developmental psychology and education suggest that young children show more advanced scientific thinking than previously thought; however, further development is needed to give rise to mature science knowledge and scientific reasoning. This complex process of development involves interactions between domain-general and domain-specific cognitive processes and abilities (such as epistemological understandings, theory of mind, metacognition, self-regulation) and environmental inputs (such as informal interactions at home and museums, and formal education and instruction), which are just beginning to be understood. The articles included in this Research Topic provide a sample of the range of current research that investigates these complex interactions and provide a road map to the development of scientific thinking that is helpful and illuminating for both theory and practice.

An important task for future research will be to determine how the basic abilities discussed in the articles included in this Research Topic come together with domain-general and domain-specific cognitive processes to influence mature scientific thinking. Although there is evidence to suggest a certain degree

of stability in scientific reasoning from middle childhood to adulthood (Bullock et al., 2009), more work is needed to connect mature scientific thinking with the basic science abilities present in younger children. Studies of this sort will be challenging given their longitudinal nature and the need for designing more appropriate measures of broad scientific reasoning abilities [see Koerber and Osterhaus (2019), van de Sande et al. (2019), and Osterhaus et al. (2020); for recent examples]. However, this work is important both from a theoretical perspective and in order to speak to the long-term value of efforts to foster scientific thinking in early education (Klahr et al., 2011).

In addition to documenting that young children can, in principle, explore science concepts systematically, the current Research Topic also speaks to various ways in which children's scientific thinking can be scaffolded. In particular, several studies show that young children need guidance in their systematic exploration of evidence, which is true across diverse contexts ranging from the classroom, to the science museum, to children's homes. Accordingly, designers of learning environments for young children need to consider both the type and amount of support that is needed by children of different ages and cognitive abilities. Offering too little support may result in children's failure to learn; in contrast, offering inappropriate or too much structure may diminish children's curiosity. Moving forward, it will be important to bring together researchers in developmental psychology and education–as is done in this Research Topic–to create effective opportunities to foster the critical 21st century skill of scientific thinking in the youngest members of society.

AUTHOR CONTRIBUTIONS

All editors contributed in substantial ways to the conception of the topic as well as in overseeing the reviewing process and to a genuinely collaborative writing of the editorial. CO and AB played an especially key role in the process. AN played a key role in bringing this team of editors together and overseeing the process. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The editors thank all the authors and reviewers for their diligent work in finalizing this Research Topic in a timely fashion.

REFERENCES

Bullock, M., Sodian, B., and Koerber, S. (2009). "Doing experiments and understanding science: development of scientific reasoning from childhood to adulthood," in *Human Development From Early Childhood to Early Adulthood: Findings From a 20 Year Longitudinal Study* (New York, NY: Psychology Press), 173–197.

Bullock, M., and Ziegler, A. (1999). "Scientific reasoning: developmental and individual differences," in *Individual development from 3 to 12: Findings from the Munich Longitudinal Study,* eds F. E. Weinert and W. Schneider (Cambridge: Cambridge University Press), 38-54.

Croker, S., and Buchanan, H. (2011). Scientific reasoning in a real-world context: The effect of prior belief and outcome on children's hypothesis-testing strategies. *Br J Dev Psychol.* 29, 409–424. doi: 10.1348/026151010X496906

Kind, P., and Osborne, J. (2017). Styles of scientific reasoning: a cultural rationale for science education? *Sci. Educ.* 101, 8–31. doi: 10.1002/sce.21251

Klahr, D., Zimmerman, C., and Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science* 333, 971–975. doi: 10.1126/science.1204528

Koerber, S., and Osterhaus, C. (2019). Individual differences in early scientific thinking: assessment, cognitive influences, and their relevance for science learning. *J. Cogn. Dev.* 20, 1–24. doi: 10.1080/15248372.2019.1620232

Köksal-Tuncer, Ö., and Sodian, B. (2018). The development of scientific reasoning: hypothesis testing and argumentation from evidence in young children. *Cogn. Dev.* 48, 135–145. doi: 10.1016/j.cogdev.2018.06.011

Kuhn, D. (2002). "What is scientific thinking and how does it develop?," in *Blackwell Handbook of Childhood Cognitive Development,* ed U. Goswami (Malden, MA: Blackwell Publishers), 371–393. doi: 10.1002/9780470996652.ch17

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., White, S. H., Klahr, D., et al. (1995). Strategies of knowledge acquisition. *Monogr. Soc. Res. Child Dev.* 60, i—157. doi: 10.2307/1166059

Kuhn, D., Iordanou, K., Pease, M., and Wirkala, C. (2008). Beyond control of variables: What needs to develop to achieve skilled scientific thinking? *Cogn. Dev.* 23, 435–451. doi: 10.1016/j.cogdev.2008.09.006

Lazonder, A. W., and Harmsen, R. (2016). Meta-analysis of inquiry-based learning: effects of guidance. *Rev. Educ. Res.* 86, 681–718. doi: 10.3102/0034654315627366

Lehrer, R., and Schauble, L. (2007). "Scientific thinking and science literacy," in *Handbook of Child Psychology,* eds W. Damon, R. M. Lerner, K. A. Renninger, and I. E. Siegel (Hoboken, NJ: Wiley), 153–196. doi: 10.1002/9780470147658.chpsy0405

Osterhaus, C., Koerber, S., and Sodian, B. (2017). Scientific thinking in elementary school: children's social cognition and their epistemological understanding promote experimentation skills. *Dev. Psychol.* 53, 450–462. doi: 10.1037/dev0000260

Osterhaus, C., Koerber, S., and Sodian, B. (2020). The Science-P Reasoning Inventory (SPR-I): Measuring emerging scientific-reasoning skills in primary school. *Int. J. Sci. Educ.* 42, 1087–1107. doi: 10.1080/09500693.2020.1748251

Ruffman, T., Perner, J., Olson, D. R., and Doherty, M. (1993). Reflecting on scientific thinking: children's understanding of the hypothesis?evidence relation. *Child Dev.* 64, 1617–1636.

Sadler, T. D. (2004). Informal reasoning regarding socioscientific issues: A critical review of research. *J. Res. Sci. Teach.* 41, 513–536. doi: 10.1002/tea.20009

Shtulman, A., and Walker, C. (2020). Developing an understanding of science. *Annu. Rev. Dev. Psychol.* 2, 111–132. doi: 10.1146/annurev-devpsych-060320-092346

Sodian, B., Zaitchik, D., and Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Dev.* 62, 753–766. doi: 10.2307/1131175

Trilling, B., and Fadel, C. (2009). *21st Century Skills: Learning for Life in Our Times*. San Francisco, CA: Jossey-Bass.

Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Dev.* 51, 1–10.

van de Sande, E., Kleemans, T., Verhoeven, L., and Segers, E. (2019). The linguistic nature of children's scientific reasoning. *Learn. Instr.* 62, 20–26. doi: 10.1016/j.learninstruc.2019.02.002

van der Graaf, J., Segers, E., and Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: dynamic assessment of the control of variables strategy. *Instr. Sci.* 43, 381–400. doi: 10.1007/s11251-015-9344-y

Vosniadou, S. (2002). "On the nature of naive physics," in *Reconsidering Conceptual Change: Issues in Theory and Practice,* eds M. Limon and L. Mason (Dordrecht: Springer), 61–76. doi: 10.1007/0-306-47637-1_3

Vosniadou, S. (2019). The development of students' understanding of science. *Front. Educ.* 4:32. doi: 10.3389/feduc.2019.00032

Vosniadou, S., and Brewer, W. F. (1992). Mental models of the earth: a study of conceptual change in childhood. *Cogn. Psychol.* 24, 535–585. doi: 10.1016/0010-0285(92)90018-W

Wason, P. C. (1968). Reasoning about a rule. *Q. J. Exp. Psychol.* 20, 273–281. doi: 10.1080/14640746808400161

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Osterhaus, Brandone, Vosniadou and Nicolopoulou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 14 May 2020
doi: 10.3389/fpsyg.2020.00904

Individual Differences in Children's Development of Scientific Reasoning Through Inquiry-Based Instruction: Who Needs Additional Guidance?

Erika Schlatter, Inge Molenaar and Ard W. Lazonder*

Behavioural Science Institute, Radboud University, Nijmegen, Netherlands

OPEN ACCESS

Edited by:
Christopher Osterhaus, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Antti Tapio Lehtinen, University of Jyväskylä, Finland
Amy Masnick, Hofstra University, United States
Özgün Köksal, Ludwig Maximilian University of Munich, Germany

***Correspondence:**
Erika Schlatter
e.schlatter@pwo.ru.nl

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *31 January 2020*
Accepted: *14 April 2020*
Published: *14 May 2020*

Citation:
Schlatter E, Molenaar I and Lazonder AW (2020) Individual Differences in Children's Development of Scientific Reasoning Through Inquiry-Based Instruction: Who Needs Additional Guidance? Front. Psychol. 11:904. doi: 10.3389/fpsyg.2020.00904

Scientific reasoning involves a person's ability to think and act in ways that help advance their understanding of the natural world. Young children are naturally inclined to engage in scientific reasoning and display an emerging competence in the component skills of, for example, hypothesizing, experimenting and evaluating evidence. Developmental psychology research has shown that same-age children often differ considerably in their proficiency to perform these skills. Part of this variation comes from individual differences in cognition; another part is due to the fact that the component skills of scientific reasoning emerge at a different age and mature at a different pace. Significantly less attention has been paid to children's capacity to improve in scientific reasoning through instruction and deliberate practice. Although elementary science lessons are generally effective to raise the skill level of a group of learners, not all children benefit equally from the instructional treatment they receive. Knowing what causes this differential effectiveness is important as it can inform the design of adaptive instruction and support. The present study therefore aimed to identify and explain how fifth-graders (N = 138) improve their scientific reasoning skills over the course of a 5-week inquiry-based physics unit. In line with our expectations, significant progress was observed in children's achievements on a written scientific reasoning test, which was administered prior to and after the lessons, as well as in their responses to the questions and assignments that appeared on the worksheets they filled out during each lesson. Children's reading comprehension and mathematical skillfulness explained a portion of the variance in children's pretest-posttest gain. As these overall results did not apply equally to all component skills of scientific reasoning, we recommend science teachers to adapt their lessons based on children's past performance in reading and math *and* their actual performance of each scientific reasoning skill. The orchestration and relative effectiveness of both adaptive science teaching approaches is an interesting topic for future research.

Keywords: scientific reasoning, inquiry-based learning, elementary education, individual differences, instructional guidance

INTRODUCTION

Elementary science education acquaints children with fundamental science concepts such as buoyancy, motion and electricity, and introduces them to the basics of doing scientific research. School science lessons make ample use of inquiry-based teaching methods, which enable children to learn to think and act in ways that help advance their understanding of the natural world (Kind and Osborne, 2017). This ability is commonly referred to as *scientific reasoning* and involves the skills of hypothesizing, experimenting and evaluating evidence (Klahr and Dunbar, 1988; Zimmerman, 2007; Kind, 2013). The main purpose of this study was to investigate how these skills develop during an inquiry-based science unit, and which cognitive characteristics predict children's level of skillfulness at the end of the unit.

The teaching and learning of scientific reasoning is a challenging task for both teachers and children. One complicating factor is that considerable individual variation exists among children in the same classroom (Koerber et al., 2015; Lazonder et al., submitted). To complicate matters further, the component skills of scientific reasoning are known to emerge at different ages and develop at a different pace (Piekny and Maehler, 2013). Kindergartners already show some initial proficiency in basic experimentation and evidence evaluation skills whereas the more difficult skill of hypothesizing usually starts developing around the age of 12. These accumulating differences point to a clear need for adaptive instruction, but until now few evidence-based guidelines for designing and delivering adaptive and age-appropriate science lessons seem to exist.

In working toward establishing such guidelines, the present study sought to unveil whether and to what extent the progress monitoring data available in schools can help predict differences in children's ability to learn scientific reasoning. Many schools have access to rich data records that portray children's developmental trajectories in the foundation skills of language and math. As these skills are related to children's scientific reasoning *performance* (e.g., Tajudin and Chinnappan, 2015; Van de Sande et al., 2019), it seems worthwhile to investigate their predictive powers for the *development* of scientific reasoning in an instructional setting. Additionally, process data collected during the lessons was analyzed in order to identify key learning moments. The insights that result from these investigations can help teachers to respond adequately to individual differences during their science lessons.

THEORETICAL FRAMEWORK

Development of Scientific Reasoning

Scientific reasoning is a multidimensional process that consists of several component skills. Although scholars diverge on the definition and labeling of these skills (for an overview, see Pedaste et al., 2015), consensus seems to exist on the core skills of hypothesizing, experimenting and evaluating evidence (Zimmerman, 2007; Kind, 2013). Even though these are difficult skills even for adults (Zimmerman, 2007), children at the preschool age already show some emerging proficiency in an skills (Sodian et al., 1991; Piekny et al., 2014; Van der Graaf et al., 2015; Köksal-Tuncer and Sodian, 2018; Van der Graaf et al., 2015) that develops steadily but slowly during the elementary school years (Kuhn, 2002; Piekny and Maehler, 2013; Koerber et al., 2015).

Although developmental growth occurs in all component skills, their emergence and pace of development varies. Experimenting is relatively easy to learn and even young children can be rather proficient in the basics of experimentation (Cook et al., 2011; Van der Graaf et al., 2018; Schalk et al., 2019; Van der Graaf et al., 2018). Hypothesizing is more difficult for children to learn (Piekny and Maehler, 2013; Schlatter et al., 2019) and this skill generally emerges late and develops slowly. Results regarding evidence evaluation are mixed. First-graders can already draw correct conclusions from perfectly covarying data (Koerber et al., 2005; Piekny and Maehler, 2013; Van der Graaf et al., 2018), but the evaluation of non-perfect covarying evidence in light of hypotheses remains difficult throughout elementary school (Piekny and Maehler, 2013).

In addition to this variation across component skills, same-age children are not equally well versed in scientific reasoning either. In a large-scale cross-sectional study using written tests in grades 2 to 4, Koerber et al. (2015) distinguished between naïve, intermediate, and advanced conceptions of scientific reasoning. Although older children generally had a more sophisticated view, all three proficiency levels were present in all participating grade levels. The cross-sectional results of Piekny and Maehler (2013) further suggest that these inter-individual differences increase with age in all component skills. For example, both the means and standard deviations of hypothesizing were low in Kindergarten, but increased from grade 1 onward. These findings indicate that, although children's hypothesizing skills undergo a steady growth, the variation among peers grows accordingly. Thus, children improve in scientific reasoning with age, but not all children improve at the same pace. Acknowledging these individual differences alongside the dissimilar difficulty levels of the component skills is vital for good science education.

Predictors of Scientific Reasoning

Several studies have examined what accounts for observed differences in children's scientific reasoning (Siler et al., 2010; Mayer et al., 2014; Wagensveld et al., 2014; Koerber et al., 2015; Wagensveld et al., 2014). Reading comprehension was a significant predictor in all these studies, whereas cognitive characteristics such as spatial reasoning, problem solving skill, and general intelligence had a less prominent and less consistent impact. Although mathematical skillfulness has been shown to correlate with scientific reasoning (Bullock and Ziegler, 1999; Koerber and Osterhaus, 2019), to the best of our knowledge, no studies have examined whether mathematical skillfulness predicts scientific reasoning. This seems remarkable because scientific reasoning tasks often require children to handle numerical data (Kanari and Millar, 2004), which, in turn, could be the reason why national curriculum agencies consider mathematical skillfulness as a prerequisite for scientific reasoning instruction (e.g., van Graft et al., 2018; Wong, 2019).

Research into the predictors of scientific reasoning either treated scientific reasoning as a unitary construct or focused on one of its components, with experimenting being the most widely studied skill. Studies that assess and report children's performance on multiple component skills are clearly underrepresented in the literature and, as a consequence, little is known about how well reading comprehension and mathematical skillfulness predict children's proficiency in separate scientific reasoning skills. Initial evidence suggests that both predictors may have differential effects. Schlatter et al. (2019) established that reading comprehension predicts performance on all subskills except hypothesizing, whereas Van de Sande et al. (2019), who administered a written test, found a strong explanatory effect of reading comprehension on this skill and lower impacts on experimenting and drawing conclusions. Osterhaus et al. (2017) found no effect of language abilities on experimenting–although it did influence children's understanding of the nature of science. These findings, although apparently contradictory, emphasize the importance of analyses at the subskill level. However, as these studies examined children's scientific reasoning *performance*, research still has to determine whether and to what extent reading comprehension and mathematical skillfulness affect and predict children's *learning* of the component skills of scientific reasoning in regular science classrooms.

In-School Learning of Scientific Reasoning

Studies examining the development of scientific reasoning in an instructional setting predominantly target children's ability to design and conduct controlled experiments. The natural development of this skill can be boosted in a short period of time through various instructional methods that often yield long-term effects. Implicit methods such as giving hints to focus the investigation on a single variable (Kuhn and Dean, 2005), dividing the research question in single-variable subquestions (Lazonder and Wiskerke-Drost, 2015), providing scaffolds (van Riesen et al., 2018) or opportunities for sustained practice (Schalk et al., 2019) all improve children's experimentation skills. Explicit instructional methods that explain and/or demonstrate the design of controlled experiments have similar benefits (Chen and Klahr, 1999; Lorch et al., 2017). A recent meta-analysis substantiated that implicit and explicit methods are equally effective for promoting experimenting skills (Schwichow et al., 2016).

The skills of hypothesizing and evaluating evidence have less often been trained in isolation, but are included in integrated studies of scientific reasoning, often using microgenetic designs. In a 3-year longitudinal study, Kuhn and Pease (2008) found that repeated practice alone promotes children's evidence evaluation skills throughout grades 4 to 6. Hypothesizing skills improved only when children were in sixth grade–despite frequent opportunities for practice in the preceding years–and individual change patterns in both skills varied considerably, with relapses to old, less-effective routines. More explicit instructional support can accelerate children's natural pace of development. Greven and Letschert (2006) showed that sixth-graders who merely investigated a multivariable system did improve their ability to evaluate evidence over the course of a 5-week inquiry-based lesson series. However, significantly higher learning gains were observed in children who received additional prediction practice exercises (that focused their attention on integrating the impact of multiple variables) or explicit instruction on the concept of multivariable causality.

To conclude, the cited studies exemplify that even short instructional interventions can promote children's scientific reasoning. Prolonged opportunities for practice have similar beneficial effects but seem more difficult to realize in regular science classrooms. Striking the right balance between independent practice and instructional guidance thus seems a major challenge elementary science teachers have to meet. This orchestration of instructional support is complicated further by the substantial variation across the component skills and among same-aged children. As a large share of this variance remains unexplained, the present study aimed to describe and explain children's development of scientific reasoning skills in inquiry-based classrooms.

Research Questions and Hypotheses

Previous research has shown that the component skills of scientific reasoning are not equally well developed and learned in upper-elementary science classes. Although these individual differences are explained in part by children's cognitive characteristics, with reading comprehension being the most robust predictor, questions remain as to how the core scientific reasoning skills of hypothesizing, experimenting and evaluating evidence develop in an instructional setting, and how developmental differences can be adequately accommodated by elementary science teachers. The present study therefore aimed to find out:

(1) To what extent fifth-graders improve their scientific reasoning skills during a 5-week inquiry-based unit;
(2) Whether observed differences in learning gains are contingent on children's reading comprehension and mathematical skillfulness; and
(3) Whether there are any key moments during this lesson series where children make marked progress in their application of the component scientific reasoning skills.

These research questions were examined in a sample of Dutch fifth-graders, who engaged in 5 weekly science lessons. Each lesson revolved around a hands-on investigation that enabled children to practice the component skills of hypothesizing, experimenting and evaluating evidence. Children' investigations were guided by worksheets and a whole-class introduction to the steps of the inquiry cycle. Learning gains were assessed by a written scientific reasoning pre-test and post-test. Learning process data were collected from the children's worksheets, whereas children's scores on standardized progress monitoring tests of reading comprehension and mathematics were obtained from the schools' administration.

Hypotheses regarding the first research question predicted that children would make progress in all scientific reasoning

skills–but not to the same degree. As previous research has shown that hypothesizing and evaluating evidence is largely beyond fifth-graders' reach, these skills were expected to improve marginally and comparably in just five lessons. Experimenting on the other hand is known to be relatively easy so children might already be rather adept in this skill and, hence, have less opportunity for improvement compared to the other skills. However, in absence of a national science curriculum and instigated by recent policy measures, many Dutch elementary schools are just beginning to systematically incorporate science in their curriculum (Inspectie van het Onderwijs, 2017), a rival hypothesis therefore predicted that children's experimentation skills are initially lower than expected based on international benchmarking studies, but will improve more rapidly over the course of the five lessons than the other component skills – a result more often observed in intervention studies (Lorch et al., 2014; Peteranderl, 2019).

The second set of hypotheses related to the prediction of learning progress. Even when no overall learning gain is found, part of the sample could have made significant progress. To explain such possibly differential progress, two predictor variables were used: reading comprehension and mathematical skillfulness. Previous studies have shown that the former consistently predicts individual differences in scientific reasoning performance. We therefore felt it safe to assume that reading comprehension would explain learning gains in all three component skills. Evidence regarding the impact of children's mathematical skillfulness is limited, but existing studies suggest that 'being good with numbers' serves as an advantage when interpreting the numerical outcomes of science experiments (Bullock and Ziegler, 1999; Koerber and Osterhaus, 2019). Children's mathematical skillfulness was therefore expected to predict learning gains in evidence evaluation.

Thirdly, children's worksheets were scrutinized for evidence of possible growth spurts in children's learning of the three component scientific reasoning skills. In absence of any theoretical and empirical underpinnings, no explicit hypothesis was made regarding the outcome of this analysis.

MATERIALS AND METHODS

This study was carried out in accordance with the recommendations of the ethics code for research with human participants in the social and behavioral sciences, as agreed upon by the Deans of Social Sciences in the Netherlands. The protocol was approved by the Ethics Committee of the Faculty of Social Sciences at Radboud University, under number 2018-074R1. Descriptive data (gender and year of birth) were collected anonymously while other data (pre- and post-test, worksheets and standardized test scores) were pseudonymized.

Participants

In the Fall of 2018, eight fifth-grade classes (in Dutch: 'groep 7') from six schools in the central and northern part of the Netherlands participated in this study. All children in these classrooms received five 1-hr lessons as part of their regular science curriculum. Passive parental consent was sought with the exception of one school, whose principal preferred active parental permission for participation. Children with parental consent (N = 154) also took a scientific reasoning pre- and post-test; the worksheets they filled out during the lessons were collected for analysis, and their progress monitoring scores on standardized tests of reading comprehension and mathematics were obtained from the school. Sixteen children were excluded from the analyses, either because they missed more than one lesson, had not taken the pre- or post-test, or because their reading and math progress monitoring records could somehow not be obtained. The final sample thus consisted of 138 participants (55% boys) who were between 8 and 12 years of age; the majority of the sample was 10 years old.

Materials

Lesson Materials

Children engaged in five science lessons that addressed elementary-school physics topics (see **Figure 1**) through an inquiry-based teaching approach, taught by the first author. All lessons were structured similarly and contained two types of activities: whole-class discussion and small-group work. Each lesson started with a plenary introduction (lesson 1) or refresher (lessons 2–5) of the inquiry cycle, and introduced children to the topic of inquiry. Children then started their first inquiry, which they completed in 20 min. In order to mimic authentic classroom practice, children conducted their investigation in dyads, which they formed themselves on an ad-hoc basis. As children chose their learning partners based on friendship rather than academic achievement and partnerships rotated during the lesson series, the chances of any systematic bias due to group formation were assumed to be negligible. The first inquiry was wrapped up during a short whole-class discussion that addressed questions such as 'who found an answer to the research question?' and 'who found a different result than hypothesized?' After the second 20-min inquiry cycle, children reconvened for a final whole-class discussion of the outcomes of the inquiry and the underlying physics principles.

The lessons were designed to practice four scientific reasoning skills: hypothesizing, experimenting, interpreting data, and drawing conclusions. Each lesson centered around a different subject-specific topic (see **Figure 1**) that children could learn about through experimentation. All experiments had three dichotomous input variables and one continuous output variable. For example, the pendulum swing experiment enabled children to manipulate the length of the rope (long or short), the weight of the pendulum (heavy or light), and the amplitude (far or close). Children used a stopwatch to measure the time it took to make five swings. In a typical lesson, the experimental equipment would be used during two inquiry sessions that were structured according to the inquiry cycle and enabled children to investigate two distinct research questions.

All inquiry sessions were supported by worksheets (see **Supplementary Appendix 1**) that assisted children in performing the four scientific reasoning skills of the inquiry cycle. This guidance consisted of a pre-specified research question and

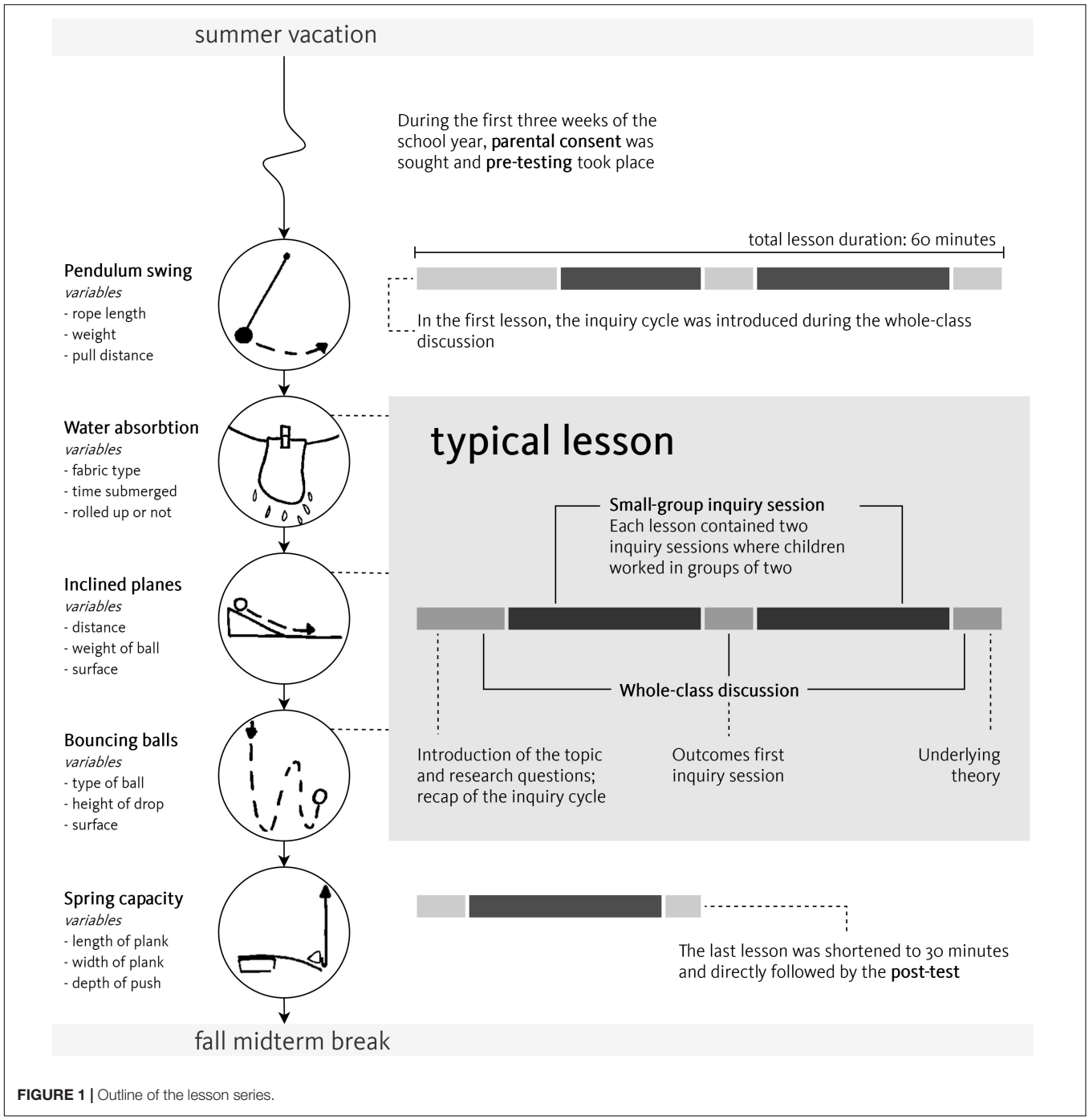


FIGURE 1 | Outline of the lesson series.

several scaffolds, that structured children's inquiry without explicitly instructing them what to do and why. Specifically, children could complete sentence starters to make their hypotheses and conclusions, and complete pre-structured tables to set up an experiment and interpret the results. The worksheets contained text and pictures that served to remind children of the research question and the variables under investigation (for an overview of inquiry topics and variables, see **Figure 1**).

This amount of support, defined by Bell et al. (2005) as *guided inquiry*, purposefully constrained the number of strategies children could apply, and has been shown to facilitate the learning of scientific reasoning skills (e.g., van Riesen et al., 2018). For example, providing a research question minimized the risk of children conceiving a research question that could not be investigated, while still providing them with a fair degree of autonomy in their inquiry. The worksheets thus had a dual purpose: in addition to being a supportive device, they served

as a measure of children's progress in scientific reasoning. Even though children conducted their investigations in dyads, they wrote down what they themselves thought to be the best hypothesis, experiment, interpretation of the data they gathered, and conclusion. As such, this process data could be used to identify where additional support was needed, and thus inform future research on adaptive science instruction.

Supplementary to the worksheets, more elaborate support was given during whole-class discussions and to individual children who indicated they were struggling with the assignments. Children who struggled were first prompted to write down what they thought was best. If they were still hesitant to work on the worksheet, guidance was slowly increased following the protocol in **Supplementary Appendix 2**. In practice, children rarely asked for help and no child asked help repeatedly for the same component skill. During the whole-class discussions, children were invited to share what they remembered about the inquiry cycle, what they found out during their investigations and what they thought were the underlying scientific principles. If answers were limited (e.g., 'we found that it made a difference'), children were encouraged to provide more detail (e.g., 'can you explain more precisely what you found?').

Scientific Reasoning Inventory

Children's scientific reasoning skills were assessed at pre- and post-test using the Scientific Reasoning Inventory (SRI; Van de Sande et al., 2019), a pencil-and-paper test consisting of 24 multiple-choice items with three to four answer options each. Items were thematically embedded in five cover stories that were meaningful and appealing to children, such as the living conditions of wildlife and sports activities.

During the original validation of the SRI, three scales emerged: hypothesis validation (which included data interpretation), experimentation and drawing conclusions (Van de Sande et al., 2019). Confirmatory factor analysis was performed and results, including the comparative fit index (CFI), root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR) are reported below. In our pre-test data, a single-factor solution had a rather poor fit, $\chi^2(252) = 437.04$, $p < 0.001$, CFI = 0.608, RMSEA = 0.067, SRMR = 0.081. The original three-factor model had a better fit, $\chi^2(249) = 354.99$, $p < 0.001$, CFI = 0.776, RMSEA = 0.051, SRMR = 0.075, and the four-factor model, with data interpretation as a separate factor, yielded comparable fit statistics, $\chi^2(246) = 352.15$, $p < 0.001$, CFI = 0.775, RMSEA = 0.051, SRMR = 0.074. While the improvement from the single-factor model to the three-factor model was significant, $\chi^2_{diff}(3) = 82.046$, $p < 0.001$, the improvement from the three-factor model to the four-factor model was not, $\chi^2_{diff}(3) = 2.841$, $p = 0.417$. We therefore decided to use the original three scales in the analyses. As a consequence, there was no one-on-one match between the SRI-scales and the skills addressed by the worksheets. Specifically, hypothesizing and interpreting outcomes were separate skills on the worksheets but combined in one SRI-scale, which we labeled 'hypothesis-evidence coordination.'

This *hypothesis-evidence coordination* scale (9 items, $\alpha_{pretest} = 0.66$, $\alpha_{posttest} = 0.74$), consisted of two types of items. Five items presented children with four research questions, and asked them to select the question that best matched the research purpose described in the cover story. The nature of these items closely resembled the way in which the skill of hypothesizing was addressed during the lessons. Four additional items measured children's ability to interpret a table with research data. These questions related to the skill of interpreting data as was addressed during the lessons. Although these nine items loaded on the same scale in the SRI, they were practiced separately during the intervention because they took place in a different stage of the inquiry cycle.

The second scale, *experimenting* (7 items, $\alpha_{pretest} = 0.47$, $\alpha_{posttest} = 0.81$), required children to select the best experiment based on the cover story. Each item presented children with three experimental designs with either two variables (2 items) or three variables (5 items). For each experiment only one experimental setup allowed for valid causal conclusions. The other experiments were either confounded, did not change any variables, or were controlled but did not manipulate the target variable.

Items on the third scale, *drawing conclusions* (8 items, $\alpha_{pretest} = 0.64$, $\alpha_{posttest} = 0.77$), contained two premises and a question about those premises children could answer with 'yes,' 'no,' or 'maybe.' These syllogisms were embedded in the overarching cover story. For example, one of the syllogisms in the sports storyline was: 'All children who will go rowing, are wearing shorts. Anna will go rowing. Is she wearing shorts?'

Reading Comprehension and Mathematical Ability

Most schools in the Netherlands participate in the student monitoring program of the National Institute for Educational Testing and Assessment [Stichting Cito Instituut voor Toetsontwikkeling]. This program includes standardized assessments of children's cognitive abilities, which are administered twice a year. The tests of reading comprehension and mathematical skillfulness were used in the present study.

The reading comprehension test provided children with different types of texts, such as short stories, newspaper articles, advertisements and instructional manuals (Weekers et al., 2011). 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 meant or choose an appropriate continuation of a story. The mathematics test had children solve 96 multiple-choice and open-ended problems that were presented either with or without context (Hop et al., 2017). Contextualized problems consisted of a short text in which the problem was outlined and a supporting picture. In problems without context, children would only be presented with the numerical operations. Sample items of both tests can be found in **Supplementary Appendix 3**.

The monitoring program provides raw scores as well as a proficiency score (I-V, with I being the highest level and V the lowest). The latter can be used to meaningfully compare scores across different versions of the monitoring program. Because all participating schools used the same student monitoring program, but not all schools used the same version, these proficiency scores were used as predictor variables. As such, the association between children's scientific reasoning and their proficiency in reading comprehension and mathematics could be

assessed without burdening children with more tests. In order to improve interpretability, proficiency scores were recoded so that 1 represented the lowest proficiency and 5 the highest proficiency.

Worksheet Scoring

The worksheets served a dual purpose in this study. In addition to being a supportive device, they were used as a process measure of children's learning. To this end, the worksheets of all five lessons were made as comparable as possible, differing only with regard to subject content (i.e., names of variables and images directly related to the subject-specific content). The questions and scaffolds were identical throughout the lesson series.

Worksheets were coded for each component skill (i.e., hypothesizing, experimenting, interpreting data, drawing conclusions). For each skill a maximum of 3 points was awarded, resulting in a maximum of 12 points per worksheet (see **Table 1** for the coding scheme). *Hypotheses* were classified according to their level of specificity using the criteria proposed by Lazonder et al. (2010). Given the young age group in the current study, the definition of a fully-specified hypothesis was slightly altered: it included the variables involved and a prediction of the direction of effect. *Experimenting* was assessed from children's use of the control-of-variables strategy (CVS; Chen and Klahr, 1999). It is important to note that this was not an all-or-nothing evaluation: even if the CVS was not applied, some points could still be awarded depending on the severity of the misconceptions (Peteranderl, 2019). At the very least, children had to understand the need for contrast, so a confounded experiment still received one point, whereas an experiment in which no variables were changed received zero points. The worksheet assignment for *interpreting data* consisted of two parts. The first part was a yes/no question that asked children whether they had observed a difference in outcomes between the two values of the focal variable. If the inference matched their data, one point was awarded. This inference should ideally be made based on multiple iterations of the same experiment. However, data gathered by children can be complex and messy (Kanari and Millar, 2004) and if this was the case, the single comparison was evaluated as a check. In the second part, children were asked to justify their inference. Two more points were awarded if children stated that they used the data to make this inference (a verbal statement of (non)covariation; Moritz in Ben-Zvi and Garfield, 2004) and/or explained what caused the result they found. *Conclusions* were, like hypotheses, evaluated in terms of their specificity (Lazonder et al., 2010). In addition to the criteria described above, the effect children mentioned in their conclusion had to match the data they gathered.

A set of 86 randomly selected worksheets was coded by a second independent rater; the intraclass correlation (ICC) was calculated as a measure of interrater reliability. The ICC was high for all component skills: hypothesizing (0.91, $p < 0.001$), experimenting (0.82, $p < 0.001$), interpreting data (0.94, $p < 0.001$), and drawing conclusions (0.89, $p < 0.001$). Differences in interrater agreement were resolved through discussion. If children were present during all lessons, nine worksheets would be available. In practice, some children missed one lesson and some worksheets got lost in the classroom. As a result, between six and nine worksheets were available per child.

Procedure

The study was carried out over a period of 6 weeks according to the setup outlined in **Figure 1**. During the 1st week, all children made the pre-test in a whole-class test setting. In weeks 2–6, children participated in five 1-hr lessons taught by the first author. Due to time constraints, the final lesson included the post-test and, hence, contained only one small-group inquiry. As the study did not aim to compare different instructional treatments, all children received the exact same lessons.

TABLE 1 | Coding scheme.

Skill	Evaluation criteria	Example
Hypothesizing	- An **effect** was described - The **direction** of the effect was described - The **variables involved** were described	'I think it makes a difference' (1 point: effect described) 'I think the surface matters for the number of bounces (2 points: effect and variables described; no direction) 'I think there will be more bounces on a hard surface' (3 points)
Experimenting	- **Comparison** is possible: at least one variable has been changed - **Fair comparison** is possible: only one variable has been changed - Experiment **aligns with the research question:** focal variable has been changed	Confounded experiment (1 point: comparison possible) Controlled experiment on non-focal variable (2 points) Controlled experiment on focal variable (3 points)
Interpreting data	- Based on the gathered data, a correct **inference** was made - The **explanation of the inference** refers to the data or outcome variable - The data on which the inference was based are **described** or - the outcome is **explained**	Part 1: Do you see a difference in the table? **yes/no** 1 point if answer aligns with data; 0 if not Part 2: How do you know? 'the number of bounces is different' (1 point: refers to outcome variable) 'on a hard surface the ball makes 5 more bounces than on a soft surface' (2 points: describes data and refers to variable)
Drawing conclusions	- The **effect that was found** was described - The **direction** of the effect was described - All **variables involved** were described	'It makes a difference' (1 point; only if this was really found) 'The surface matters for the number of bounces' (2 points) 'The ball made more bounces on a hard surface' (3 points)

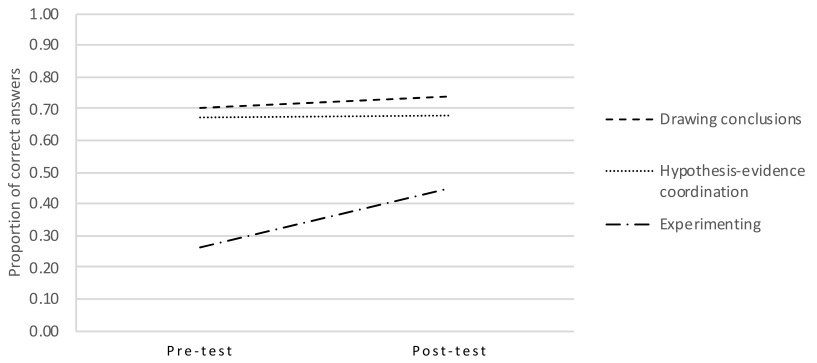


FIGURE 2 | Pre- and post-test scores per component scientific reasoning skill.

TABLE 2 | Descriptive statistics on reading comprehension and mathematics.

	Level				
	I	II	III	IV	V
Reading comprehension	30.4%	25.4%	19.6%	15.2%	9.4%
Mathematics	26.1%	17.4%	25.4%	18.8%	12.3%

TABLE 3 | Pre- and post-test scores on the Scientific Reasoning Inventory.

	Pre-test		Post-test		Gain	
	M	*SD*	*M*	*SD*	*M*	*SD*
Hypothesis-evidence coordination	0.67	0.23	0.68	0.27	0.01	0.21
Experimenting	0.26	0.22	0.45	0.35	0.20	0.38
Drawing conclusions	0.70	0.22	0.74	0.24	0.04	0.25
Overall	0.56	0.15	0.63	0.22	0.07	0.17

Scores are reported as proportion of correct answers.

RESULTS

Standardized progress monitoring data of reading comprehension and mathematics were obtained from 138 children (see **Table 2**); their pre- and post-test scores on the SRI are shown in **Figure 2** and **Table 3**. These data show that, overall, children improved in scientific reasoning, but improvement rates differed among component skills. In order to explore these differences in scores and establish their relations with reading comprehension and mathematical skillfulness, a repeated measures multivariate analysis of covariance (MANCOVA) was carried out with time and component skill as within-subject variables, and reading comprehension level and mathematics level as between-subject covariates.

Development and Prediction

Multivariate test results showed a main effect of time, Wilk's $\lambda = 0.801$, $F(1, 134) = 33.393$, $p < 0.001$ and skill, Wilk's $\lambda = 0.883$, $F(1, 134) = 8.840$, $p < 0.001$. In addition to these main effects, an interaction was found between time and skill, Wilk's $\lambda = 0.796$, $F(2, 133) = 17.014$, $p < 0.001$, indicating asynchronous development of the component skills over time. Lastly, three-way interactions were found between time, skill and reading comprehension, Wilk's $\lambda = 0.943$, $F(2, 133) = 4.040$, $p = 0.020$, and time, skill and mathematical skillfulness, Wilk's $\lambda = 0.907$, $F(2, 133) = 6.842$, $p = 0.001$, indicating that both reading comprehension and mathematical skillfulness explain variation in development of the component skills throughout the lesson series.

Both the data in **Table 3** and the significant time × skill interaction suggest that there may be subgroups of children who learned more than others. To examine this possibility, children's change in scores from pre- to post-test were visualized in density plots for each component skill (**Figure 3**). In these plots, the diagonal line stands for 'no development'; the area above the diagonal represents a decline in score, and the area below the diagonal indicates progress. For hypothesis-evidence coordination and drawing conclusions, most dots accumulate around the diagonal, meaning that children generally made little progress in these skills. A similar pattern was found for experimenting, except that there was an additional group of dots in the lower right corner. Thus, although the majority of children hardly progressed in experimenting, a small group did. It is noteworthy that the two areas are horizontally aligned. This means that some children who scored very low on the pre-test still learned to experiment very well.

In order to further explore the three-way interactions, parameter estimates were requested for pre- and post-test scores as well as for the gain scores (**Table 4**). These showed that for hypothesis-evidence coordination, both reading comprehension and mathematical skillfulness related to pre- and post-test scores. The predictors did not relate to gain scores on this skill, likely because there was very little progress. For experimenting, pre-test scores were not related to reading comprehension or mathematics, while post-test and gain scores were. Drawing conclusions was not related to children's reading comprehension or mathematical skillfulness at all.

Key Learning Moments

The third research question addressed children's learning process by identifying possible key learning moments during the lesson series. The worksheets children filled out during the lessons

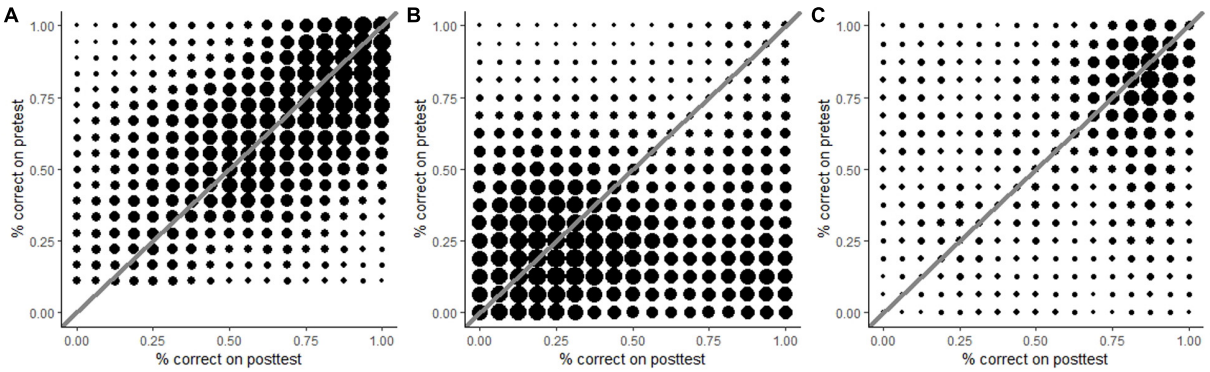


FIGURE 3 | Density plots for hypothesis-evidence coordination **(A)**, experimenting **(B)**, and drawing conclusions **(C)**. For all component skills, most scores cluster around the diagonal, indicating limited growth. For experimenting, a second cluster can be seen in the lower right corner, indicating a large improvement for a small group of children.

TABLE 4 | Parameter estimates for interaction effects.

	Pre-test		Post-test		Gain scores	
	β	*p*	β	*p*	β	*p*
Hypothesis-evidence coordination						
Reading	0.133	< 0.001	0.189	< 0.001	0.056	0.179
Math	0.088	0.001	0.089	0.003	0.001	0.976
Experimenting						
Reading	0.024	0.591	0.277	< 0.001	0.253	< 0.001
Math	0.039	0.231	0.242	< 0.001	0.202	< 0.001
Drawing conclusions						
Reading	0.021	0.603	0.062	0.165	0.041	0.423
Math	0.030	0.324	0.039	0.235	0.009	0.808

Previous analyses showed that gains in hypothesis-evidence coordination and drawing conclusions were not significant.

provided insight in this. A partial correlation between overall post-test scores (controlled for pre-test scores) and average worksheet scores was found, Spearman's $\rho = 0.408$, $p < 0.001$, warranting further inspection of the process data summarized in **Table 5**. The partial correlation coefficients in this table show that the association between post-test and worksheet was consistent for some, but not all component skills. Specifically, hypothesizing and drawing conclusions (worksheets) were not related with any of the component skills measured by the Scientific Reasoning Inventory (SRI). Experimenting (worksheets) on the other hand did correlate with experimenting (SRI) as well as with hypothesis-evidence coordination (SRI). Interpreting data (worksheets) was associated with drawing conclusions (SRI).

In addition to correlations between children's in-class performance and their achievements on the SRI, children's progress throughout the lessons was examined. First, visual inspection of the line graphs in **Figure 4** helped determine whether progress was actually made, and if so, at which moment(s) during the lesson series this growth was most pronounced. For hypothesizing, the slope appears more or less level, indicating no or very moderate improvement. Progress in the other three component skills appears to be made between the first and third lesson, after which it levels off. The first, third and fifth lesson were therefore used as anchor points in children's developmental trajectories. Repeated measures ANOVAs were performed to compare the scores on each component skill at these three timepoints. As expected based on the line graph, no main effect was found for hypothesizing, Wilk's $\lambda = 0.984$, $F(2, 115) = 0.932$, $p = 0.400$, while significant within-subject differences were found for experimenting, Wilk's $\lambda = 0.565$, $F(2, 115) = 44.287$, $p < 0.001$, interpreting data, Wilk's $\lambda = 0.659$, $F(2, 115) = 29.706$, $p < 0.001$, and drawing conclusions, Wilk's $\lambda = 0.770$, $F(2,$

TABLE 5 | Average worksheet scores and partial Spearman's rank correlations with post-test scores.

			Scientific Reasoning Inventory[1]		
			H-E coordination	Experimenting	Drawing conclusions
Worksheets	***M***	***SD***	ρ	ρ	ρ
Hypothesizing	1.40	0.57	0.122	0.145	0.136
Experimenting	1.96	0.63	0.308**	0.492**	0.120
Interpreting data	1.59	0.55	0.121	0.102	0.189*
Drawing conclusions	1.21	0.64	0.076	0.088	0.034

*Worksheet scores ranged from 0 to 3 points. [1]Post-test scores, controlled for pre-test; *p ≤ 0.05; **p ≤ 0.001.*

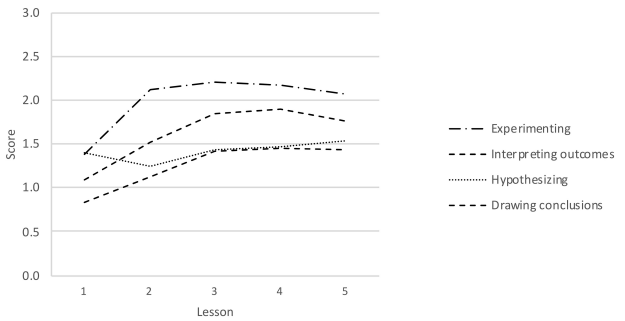


FIGURE 4 | Worksheet scores per component scientific reasoning skill in each lesson.

TABLE 6 | Key learning moments in children' scientific reasoning skills inferred from their worksheet scores.

	Lesson	*M*	*SD*	Change[1]	p_{change}
Hypothesizing	1	1.35	0.90		
	3	1.45	0.77	0.10	0.574
	5	1.49	1.10	0.04	0.978
Experimenting	1	1.37	0.80		
	3	2.19	0.88	0.82	<0.001
	5	2.07	1.03	−0.12	0.387
Interpreting data	1	1.09	0.97		
	3	1.85	0.74	0.76	<0.001
	5	1.78	0.83	−0.07	0.734
Drawing conclusions	1	0.84	1.00		
	3	1.44	0.84	0.60	<0.001
	5	1.45	1.13	0.01	1.000

Worksheet scores ranged from 0 to 4 points; [1]compared to previous.

115) = 17.145, $p < 0.001$. Pairwise comparisons across the three timepoints were made to pinpoint when learning took place. The results in **Table 6** show that for experimenting, interpreting data and drawing conclusions, significant progress was made between lessons 1 and 3, but not between lessons 3 and 5.

To assess whether improvement of the worksheet scores could be explained by children's reading comprehension and mathematical skillfulness, a 3 (lessons) × 4 (skills) MANCOVA was performed, with reading comprehension and mathematical skillfulness as covariates. Multivariate test results showed no significant three-way interaction between lesson, skill and reading comprehension, Wilk's $\lambda = 0.908$, $F(6, 109) = 1.833$, $p = 0.099$. Between lesson, skill and mathematical skillfulness a three-way interaction was found, Wilk's $\lambda = 0.881$, $F(6, 109) = 2.461$, $p = 0.029$. However, further analysis of each component skill did not yield significant interactions between time and mathematical skillfulness. Thus, although mathematical skillfulness appears to predict progress in some component skills of scientific reasoning, this effect is not large enough to detect with more specific analyses.

DISCUSSION

The main purpose of this study was to investigate how children's scientific reasoning develops during an inquiry-based science unit, and which cognitive characteristics predict progress of its component skills. Process data gathered during the lessons was analyzed to identify key moments during the lesson series when this progress was most pronounced. The findings, in short, point to a differential instructional effectiveness which should be taken into account in designing future adaptive learning arrangements.

Considerable diversity was observed in children's proficiency in and learning of scientific reasoning. Although there were significant overall gains on the SRI, this improvement did not apply equally to all component skills. Specifically, children advanced their experimenting skills, but not their ability to coordinate hypotheses with evidence and draw conclusions. Overall gains were explained by children's reading comprehension and mathematical skillfulness, as was their progress on experimenting skills and post-test performance on the hypothesis-evidence coordination items. However, both predictors explained neither progress nor proficiency in drawing conclusions. Finally, children's worksheets evidenced progress over the lessons on experimenting, interpreting data and drawing conclusions, but not on hypothesizing. Most progress was made during the first half of the lesson series. These main outcomes of the study are discussed further below.

Predicting Progress in Scientific Reasoning

The first two research questions focused on children's progress on the component skills of scientific reasoning, with reading comprehension and mathematical skillfulness as predictors. Very little to no progress was expected to occur for hypothesis-evidence coordination and drawing conclusions, which indeed turned out to be the case. Although these component skills are often deemed more difficult than experimenting, pre-test scores were rather high in the current study. Still, the complete absence of progress is somewhat remarkable and suggests that both skills are not only hard to perform but also difficult to improve. No interactions were found between the predictor variables and progress on either hypothesis-evidence coordination or drawing conclusions, but children's *proficiency* in hypothesis-evidence coordination interacted with reading comprehension and mathematical skillfulness on both pre- and post-test. This result seems understandable because the scale combined items that tap into the ability to identify appropriate research questions and interpret data, which are component skills that were expected to interact with both predictor variables.

Our hypotheses regarding experimenting were twofold: we either expected to find high pre-test scores and little progress, or low pre-test scores and substantial growth. Evidence was found for the latter hypothesis, although post-test scores for experimenting were lower than those for hypothesis-evidence coordination and drawing conclusions. This

is noteworthy because experimenting is often regarded as one of the least difficult scientific reasoning skills to learn. The large standard deviations on the post-test imply that some children had improved more than others, which was confirmed by the interactions of both the post-test scores and progress with reading comprehension and mathematical skillfulness. In combination with the density plots shown in **Figure 3**, it therefore seems plausible that some, but not all children developed adequate experimentation strategies through structured, repeated practice. Informal observations during the lessons further indicated that some children realized that the research question could not be answered based on a confounded experiment. As the worksheets did not explicitly link experimental design to drawing conclusions, conceptualizing this connection required unsupported inferencing. The significant impact of children's language and math skills suggests that only children with relatively high intellectual abilities were able to make this inference.

Progress on Scientific Reasoning During the Lessons

Children's entries on the worksheets were analyzed to unveil key moments in the learning process where marked progress in scientific reasoning was made. Notable improvements in experimenting, interpreting data and drawing conclusions occurred between lesson 1 and lesson 3, whereas no progress in hypothesizing was made over the five lessons. The latter result may be due to the fact that, unlike the other component skills, children's hypotheses were rarely addressed during the whole-class discussions. Another possibility is that hypothesizing is easier if one has a theoretical basis on the topic of inquiry (Koslowski et al., 2008), which the children in our study had not or to an insufficient degree. The lack of growth in hypothesizing skills might be attributable to a combination of these factors.

Progress on the other component skills occurred between lesson 1 and lesson 3. Interestingly, children's performance stabilized after the third lesson, despite the absence of a ceiling effect. This raises the question as to why progress leveled off before mastery was reached. A possible answer lies in the design principles underlying the lesson series. Both the lessons and the worksheets were highly structured (guided inquiry, Bell et al., 2005) but contained few explicit directions and explanations. The available implicit guidance enabled children to improve their scientific reasoning to some extent, meaning that additional growth may require additional guidance, extended practice, or both.

Using a combination of instructional support measures might help sustain children's progress beyond the third lesson. Previous research comparing open and guided inquiry to direct instruction (e.g., Alfieri et al., 2011; Wagensveld et al., 2014; Vorholzer et al., 2018; Wagensveld et al., 2014) indicated that open inquiry was often ineffective, whereas guided inquiry or direct instruction yielded higher learning outcomes. Using data from the 2015 Trends in International Mathematics and Science Study (TIMSS), Teig et al. (2018) also concluded that inquiry can be an effective approach, but only when combined with other, more explicit forms of guidance. Along these lines, more specific directions by the teacher or through the worksheets could have caused the children in our study to make significant progress on hypothesizing and to fully master the skills of experimenting, interpreting data and drawing conclusions. What these instructions should entail and how they are best combined with the scaffolding offered by the worksheets are interesting questions for future research.

Toward Adaptive Science Instruction

The present findings suggest that some children need little support to improve their scientific reasoning skills, whereas others seem to require more or more specific guidance. The worksheet data show that children improved in all scientific reasoning skills except hypothesizing; this progress was often modest and occurred in the first half of the lesson series. In order to help children further improve their scientific reasoning, we have three suggestions. First, guidance could be increased on component skills that are particularly difficult to learn, such as hypothesizing. Second, considering the relations found between scientific reasoning and children's reading comprehension and mathematical skillfulness, progress monitoring data of these school subjects can help teachers to adapt their science lessons in advance, for instance by planning to offer more or more explicit guidance to children with lower levels of reading comprehension. Third, monitoring in-class performance can inform teachers when children need additional support. Using this information to make instant adjustments above and beyond the pre-planned adaptations could be a crucial next step in the improvement of elementary science education.

Strengths and Limitations

On the positive side, this study examined multiple component skills of scientific reasoning under rather uniform conditions. As argued by Koerber and Osterhaus (2019), cross-study comparisons of proficiency and developmental growth in distinct scientific reasoning skills are likely confounded by differences in learner characteristics and task settings. Their plea for more comprehensive investigations of scientific reasoning was met here, and allows for more valid conclusions on the relative ease or difficulty with which individual scientific reasoning skills are acquired during elementary science lessons.

Another asset of this study is the use of two complementary data sources: the SRI and the worksheets. The origin of an instrument (existing or made for the study) can affect the outcomes (Schwichow et al., 2016). So in order to shed more light on children's science learning in regular classrooms, but without compromising experimental validity, we combined scores on the experimentally valid SRI, administered in a test setting, with more ecologically valid data from the worksheets children filled out during the lessons.

Although this approach yielded valuable insights in the development of some scientific reasoning skills, an unforeseen

discrepancy between these two data sources arose. Although the SRI and the worksheets both targeted the same component skills (hypothesizing, experimenting, interpreting data and drawing conclusions), factor analysis of the SRI-items in both the validation study and the current study required us to combine two of these skills in a single scale. This complicated the comparison of children's scores on the worksheets and the SRI.

This measurement inconsistency is inconvenient because different proficiency patterns emerged for the two test modalities, which are now difficult to explain. While the worksheets outcomes followed the hypothesized proficiency pattern, with highest scores for experimenting and lowest for hypothesizing, the SRI scores for hypothesis-evidence coordination and drawing conclusions were high and scores on experimenting were low. Strong claims about what accounted for these discrepancies cannot be made, but there are several possible explanations.

First, differences in test item format may have played a role. Previous studies showed that the type of data greatly influences the ease of interpretation (Kanari and Millar, 2004; Masnick and Morris, 2008). The hypothesis-evidence coordination items on the SRI featured unambiguous, dichotomous outcomes that were relatively easy to interpret, whereas the data children gathered during the lessons were continuous and more messy. Although both called upon children's ability to interpret data, requirements on the SRI were relatively limited. The high scores on hypothesis-evidence coordination and drawing conclusions suggest that the SRI taps children's basic proficiency in these component skills, whereas the worksheet provides a more authentic assessment. Secondly, surface characteristics may have limited comparability too (Stiller et al., 2016). For example, longer questions and data tables (as used on the SRI scale hypothesis-evidence coordination) can decrease difficulty, whereas the longer response options (which were used on the worksheets for hypothesizing and drawing conclusions) may increase difficulty.

Finally, reliability of the experimenting scale of the SRI pre-test was low. This was probably caused by the fact that children did not have much experience with experimenting, and because the item format was relatively difficult for them. As a result, the range of scores on the pre-test was small and this limited variability may have affected Cronbach's α .

Implications and Directions for Further Research

Although the present study provides initial directions for adaptive science education, future research is needed to assess the effectiveness of these adaptions. This and other studies show that scientific reasoning can be taught to children of all cognitive levels (Zohar and Dori, 2003), yet less is known about how the needs of individual children in a class are best met. So although our findings indicate that teachers can base instructional adaptations on children's proficiency in reading and math, research should investigate additional ways to adapt instruction in scientific reasoning.

Although the relationship between reading comprehension and scientific reasoning is well-established and caused some to conclude that scientific reasoning is linguistic in nature (Van de Sande et al., 2019), the relation between mathematical skillfulness and scientific reasoning has only recently been shown (Koerber and Osterhaus, 2019). The current study confirms that such a relationship exists. Acknowledging the impact of mathematical skillfulness is important for the effective teaching of scientific reasoning, which can be more thoughtfully designed bearing this information in mind.

CONCLUSION

Fifth-graders generally improved in scientific reasoning during a 5-week inquiry-based lesson series. They made progress in all constituent skills except hypothesizing, mainly during the first half of the lesson series, and consolidated their increased experimentation skills on the post-test. Reading comprehension and mathematical skillfulness accounted for part of the variance in children's progress and proficiency scores, and offer fertile grounds for adaptivity. However, more research is needed to fully grasp the individual variation in children's science learning and explore ways to accommodate these differences. The outcomes of these studies contribute to the design of effective elementary science education for all.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of the Faculty of Social Sciences at Radboud University (ECSS). Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

ES, IM, and AL contributed to the conception and design of the study. ES collected and analyzed the data. ES wrote the first draft of the manuscript and prepared the tables and figures. IM and AL gave feedback on all parts of the manuscript draft and rewrote several sections. All authors contributed to manuscript revision, read and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00904/full#supplementary-material

REFERENCES

Alfieri, L., Brooks, P. J., Aldrich, N. J., and Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *J. Educ. Psychol.* 103, 1–18. doi: 10.1037/a0021017

Bell, R. L., Smetana, L., and Binns, I. (2005). Simplifying inquiry instruction. *Sci. Teach.* 72, 30–33.

Ben-Zvi, D., and Garfield, J. (2004). *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. Dordrecht: Kluwer Academic Publishing.

Bullock, M., and Ziegler, A. (1999). *Scientific Reasoning: Developmental and Individual Differences Individual Development from 3 to 12: Findings from the Munich Longitudinal Study*. Cambridge, MA: Cambridge University Press, 38–54.

Chen, Z., and Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Dev.* 70, 1098–1120. doi: 10.1111/1467-8624.00081

Cook, C., Goodman, N. D., and Schulz, L. E. (2011). Where science starts: spontaneous experiments in preschoolers' exploratory play. *Cognition* 120, 341–349. doi: 10.1016/j.cognition.2011.03.003

Greven, J., and Letschert, J. (2006). *Kerndoelen Primair Onderwijs*. Netherlands: The Hague.

Hop, M., Janssen, J., and Engelen, R. (2017). *Wetenschappelijke Verantwoording Rekenen-Wiskunde 3.0 voor groep* 7. Netherlands: Arnhem.

Inspectie van het Onderwijs (2017). *Pijl.Natuur en Techniek [Level of Science Education and Performance]*. Utrecht: Inspectie van het Onderwijs [Dutch Inspectorate of Education].

Kanari, Z., and Millar, R. (2004). Reasoning from data: how students collect and interpret data in science investigations. *J. Res. Sci. Teach.* 41, 748–769. doi: 10.1002/tea.20020

Kind, P., and Osborne, J. (2017). Styles of scientific reasoning: A cultural rationale for science education? *Sci. Edu.* 101, 8–31. doi: 10.1002/sce.21251

Kind, P. M. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *J. Res. Sci. Teach.* 50, 530–560. doi: 10.1002/tea.21086

Klahr, D., and Dunbar, K. (1988). Dual space search during scientific reasoning. *Cogn. Sci.* 12, 1–48. doi: 10.1207/s15516709cog1201_1

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., and Sodian, B. (2015). The development of scientific thinking in elementary school: a comprehensive inventory. *Child Dev.* 86, 327–336. doi: 10.1111/cdev.12298

Koerber, S., and Osterhaus, C. (2019). Individual differences in early scientific thinking: assessment, cognitive influences, and their relevance for science learning. *J. Cogn. Dev.* 20, 510–533. doi: 10.1080/15248372.2019.1620232

Koerber, S., Sodian, B., Thoermer, C., and Nett, U. (2005). Scientific reasoning in young children: preschoolers' ability to evaluate covariation evidence. *Swiss J. Psychol.* 64, 141–152. doi: 10.1024/1421-0185.64.3.141

Köksal-Tuncer, Ö, and Sodian, B. (2018). The development of scientific reasoning: hypothesis testing and argumentation from evidence in young children. *Cogn. Dev.* 48, 135–145. doi: 10.1016/j.cogdev.2018.06.011

Koslowski, B., Marasia, J., Chelenza, M., and Dublin, R. (2008). Information becomes evidence when an explanation can incorporate it into a causal framework. *Cogn. Dev.* 23, 472–487. doi: 10.1016/j.cogdev.2008.09.007

Kuhn, D. (2002). "What is scientific thinking and how does it develop?," in *Blackwell Handbook of Childhood Cognitive Development*, ed. U. Goswami (Hoboken, NJ: Blackwell Publishing), 371–393.

Kuhn, D., and Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychol. Sci.* 16, 866–870. doi: 10.1111/j.1467-9280.2005.01628.x

Kuhn, D., and Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cogn. Instruct.* 26, 512–559. doi: 10.1080/07370000802391745

Lazonder, A. W., Hagemans, M. G., and de Jong, T. (2010). Offering and discovering domain information in simulation-based inquiry learning. *Learn. Instruct.* 20, 511–520. doi: 10.1016/j.learninstruc.2009.08.001

Lazonder, A. W., and Wiskerke-Drost, S. (2015). Advancing scientific reasoning in upper elementary classrooms: direct instruction versus task structuring. *J. Sci. Educ. Technol.* 24, 69–77. doi: 10.1007/s10956-014-9522-8

Lorch, R. F., Lorch, E. P., Freer, B., Calderhead, W. J., Dunlap, E., Reeder, E. C., et al. (2017). Very long-term retention of the control of variables strategy following a brief intervention. *Contemp. Educ. Psychol.* 51, 391–403. doi: 10.1016/j.cedpsych.2017.09.005

Lorch, R. F., Lorch, E. P., Freer, B. D., Dunlap, E. E., Hodell, E. C., and Calderhead, W. J. (2014). Using valid and invalid experimental designs to teach the control of variables strategy in higher and lower achieving classrooms. *J. Educ. Psychol.* 106, 18–35. doi: 10.1037/a0034375

Masnick, A., and Morris, B. J. (2008). Investigating the development of data evaluation: the role of data characteristics. *Child Dev.* 79, 1032–1048. doi: 10.1111/j.1467-8624.2008.01174.x

Mayer, D., Sodian, B., Koerber, S., and Schwippert, K. (2014). Scientific reasoning in elementary school children: assessment and relations with cognitive abilities. *Learn. Instruct.* 29, 43–55. doi: 10.1016/j.learninstruc.2013.07.005

Osterhaus, C., Koerber, S., and Sodian, B. (2017). Scientific thinking in elementary school: children's social cognition and their epistemological understanding promote experimentation skills. *Dev. Psychol.* 53, 450–462. doi: 10.1037/dev0000260

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., et al. (2015). Phases of inquiry-based learning: definitions and the inquiry cycle. *Educ. Res. Rev.* 14, 47–61. doi: 10.1016/j.edurev.2015.02.003

Peteranderl, S. (2019). *Experimentation Skills of Primary School Children*. Doctoral thesis, ETH Zürich, Switzerland.

Piekny, J., Grube, D., and Maehler, C. (2014). The development of experimentation and evidence evaluation skills at preschool age. *Int. J. Sci. Educ.* 36, 334–354. doi: 10.1080/09500693.2013.776192

Piekny, J., and Maehler, C. (2013). Scientific reasoning in early and middle childhood: the development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *Br. J. Dev. Psychol.* 31, 153–179. doi: 10.1111/j.2044-835X.2012.02082.x

Schalk, L., Edelsbrunner, P. A., Deiglmayr, A., Schumacher, R., and Stern, E. (2019). Improved application of the control-of-variables strategy as a collateral benefit of inquiry-based physics education in elementary school. *Learn. Instruct.* 59, 34–45. doi: 10.1016/j.learninstruc.2018.09.006

Schlatter, E., Lazonder, A. W., Molenaar, I., and Janssen, N. (2019). Unraveling scientific reasoning: explaining children's performance in constituent skills. *Paper Presented at the EARLI 2019*, (Aachen: EARLI).

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., and Härtig, H. (2016). Teaching the control-of-variables strategy: a meta-analysis. *Dev. Rev.* 39, 37–63. doi: 10.1016/j.dr.2015.12.001

Siler, S. A., Klahr, D., Magaro, C., Willows, K., and Mowery, D. (2010). Predictors of transfer of experimental design skills in elementary and middle school children. *Paper presented at the 10th International Conference on Intelligent Tutoring Systems*, (Pittsburgh, PA: Springer).

Sodian, B., Zaitchik, D., and Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Dev.* 62, 753–766. doi: 10.1111/j.1467-8624.1991.tb01567.x

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., et al. (2016). Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty. *Assess. Eval. High. Educ.* 41, 721–732. doi: 10.1080/02602938.2016.1164830

Tajudin, N. A. M., and Chinnappan, M. (2015). Exploring Relationship between Scientific Reasoning Skills and Mathematics Problem Solving. Mathematics Education Research Group of Australasia.

Teig, N., Scherer, R., and Nilsen, T. (2018). More isn't always better: the curvilinear relationship between inquiry-based teaching and student achievement in science. *Learn. Instruct.* 56, 20–29. doi: 10.1016/j.learninstruc.2018.02.006

Van de Sande, E., Kleemans, M., Verhoeven, L., and Segers, E. (2019). The linguistic nature of children's scientific reasoning. *Learn. Instruct.* 62, 20–26. doi: 10.1016/j.learninstruc.2019.02.002

Van der Graaf, J., Segers, E., and Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: dynamic assessment of the control of variables strategy. *Instruct. Sci.* 43, 381–400. doi: 10.1007/s11251-015-9344-y

Van der Graaf, J., Segers, E., and Verhoeven, L. (2018). Individual differences in the development of scientific thinking in kindergarten. *Learn. Instruct.* 56, 1–9. doi: 10.1016/j.learninstruc.2018.03.005

van Graft, M., Tank, M. K., Beker, T., and van der Laan, A. (2018). *Wetenschap en Technologie in Het Basis- En Speciaal Onderwijs: Richtinggevend Leerplankader Bij Het Leergebied Oriëntatie op Jezelf En De Wereld*. Enschede: SLO.

van Riesen, S., Gijlers, H., Anjewierden, A., and de Jong, T. (2018). Supporting learners' experiment design. *Educ. Technol. Res. Dev.* 66, 475–491. doi: 10.1007/s11423-017-9568-4

Vorholzer, A., von Aufschnaiter, C., and Boone, W. J. (2018). 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.* 50, 333–359. doi: 10.1007/s11165-018-9691-1

Wagensveld, B., Segers, E., Kleemans, T., and Verhoeven, L. (2014). Child predictors of learning to control variables via instruction or self-discovery. *Instruct. Sci.* 43, 365–379. doi: 10.1007/s11251-014-9334-5

Weekers, A., Groenen, I., Kleintjes, F., and Feenstra, H. (2011). *Wetenschappelijke Verantwoording Papieren Toetsen Begrijpend Lezen voor groep 7 en 8*. Arnhem: CITO.

Wong, V. (2019). 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.* 41, 1805–1826. doi: 10.1080/09500693.2019.1641249

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Dev. Rev.* 27, 172–223. doi: 10.1016/j.dr.2006.12.001

Zohar, A., and Dori, Y. (2003). Higher order thinking skills and low-achieving students: are they mutually exclusive? *J. Learn. Sci.* 12, 145–181. doi: 10.1207/S15327809JLS1202_1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Schlatter, Molenaar and Lazonder. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 26 May 2020
doi: 10.3389/fpsyg.2020.01050

Scientific Reasoning in Biology – the Impact of Domain-General and Domain-Specific Concepts on Children's Observation Competency

Janina Klemm[1], Pamela Flores[1], Beate Sodian[2] and Birgit J. Neuhaus[1]*

[1] Biology Education, Faculty of Biology, Ludwig-Maximilians-Universität München, Munich, Germany, [2] Department of Psychology and Educational Sciences, Institute for Developmental and Educational Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

Research on the development of scientific reasoning has put the main focus on children's experimentation skills, in particular on the control-of-variables strategy. However, there are more scientific methods than just experimentation. Observation is defined as an independent scientific method that includes not only the description of what is observed, but also all phases of the scientific inquiry, such as questioning, hypothesizing, testing, and interpreting. Previous research has shown that the quality of observations depends on specific knowledge in the domain. We argue that observation competency shares the domain-general ability to differentiate hypotheses from evidence with other scientific methods. The present study investigates the relations of both domain-general scientific thinking and domain-specific knowledge in biology with observation competency in grade K children. We tested relations between observation competency, domain-general scientific reasoning, domain-specific knowledge, and language abilities of 75 children (age 4;9 to 6;7). Both scientific reasoning and domain-specific knowledge proved to be significant predictors of observation competency, explaining 35% of the variance. In a mediation analysis, we found a significant indirect effect of language via these two predictors. Thus, the present results indicate that observation skills require not only domain-specific knowledge but also domain-general scientific reasoning abilities.

Keywords: scientific reasoning, domain-specific, domain-general, observation competency, kindergarten, biology

OPEN ACCESS

Edited by:
Stella Vosniadou, Flinders University, Australia

Reviewed by:
Carolyn Palmquist, Amherst College, United States
Andrew Shtulman, Occidental College, United States

***Correspondence:**
Pamela Flores
pamela.flores@biologie.uni-muenchen.de;
pamela.flores@bio.lmu.de

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *16 January 2020*
Accepted: *27 April 2020*
Published: *26 May 2020*

Citation:
Klemm J, Flores P, Sodian B and Neuhaus BJ (2020) Scientific Reasoning in Biology – the Impact of Domain-General and Domain-Specific Concepts on Children's Observation Competency. Front. Psychol. 11:1050. doi: 10.3389/fpsyg.2020.01050

INTRODUCTION

Scientific thinking in children, which is understood as "the application of the methods or principles of scientific inquiry to reasoning or problem-solving situations" (Zimmerman, 2007), has been primarily studied with respect to experimentation skills. Young children's ability to design experiments and to draw valid conclusions from data has traditionally been described as severely deficient, lacking the fundamental conceptual differentiation of hypotheses from evidence (Kuhn, 1989). However, a growing body of recent research indicates that elementary school students and

even kindergarteners may, in fact, be able to distinguish hypotheses from evidence and reason about the relation between the two in simple, knowledge-lean tasks. Sodian et al. (1991) showed that first- and second-graders were able to distinguish hypothesis testing from effect production and preferred a conclusive test for a simple hypothesis over an inconclusive one. Subsequent research by Piekny et al. (2013b) showed that even 44% of 5-year-olds were able to pass this task. More generally, kindergarteners seem to be able to choose adequate experiments (Leuchter et al., 2014; van der Graaf et al., 2015) and to interpret simple data sets (Koerber et al., 2005; Piekny et al., 2013a), unless when biased by prior beliefs (Koerber et al., 2005; Croker and Buchanan, 2011).

Once young children understand the inferential relation between hypotheses and evidence, they should be able to explore phenomena in the real world guided by their ideas (hypotheses) and to interpret data (observations) with respect to these hypotheses. To date, young children's exploration skills have been mostly studied in causal learning paradigms, in which arbitrary relations between causal factors and an effect (e.g., a lightbox) had to be discovered (e.g., Gopnik and Wellman, 2012). In the present paper, we focus on kindergarteners' exploration skills in a knowledge-rich real-world domain, the observation of animals in biology.

Observation is a key research method and an important element of science curricula (Johnston, 2009). It is relevant for social sciences (qualitative and quantitative observation of behavior) and for natural sciences, such as physics (in the field of astronomy) and biology, as it was the underlying method for Darwin's development of the theory of evolution (Kohlhauf et al., 2011). It is important to distinguish clearly between the different meanings that are ascribed to observation in the literature. It is, on one hand, regarded as a basic process in scientific research: It is needed in all stages of an inquiry and is therefore relevant for other scientific methods; e.g., when experimenting, we need to be able to make observations in the different conditions of the design. On the other hand, it is defined as an independent and complex research method that includes not only the description of what is observed, but also all phases of the scientific inquiry, such as questioning, hypothesizing, testing, and interpreting (Kohlhauf et al., 2011; see also Oguz and Yurumezoglu, 2007; Eberbach and Crowley, 2009).

From infancy, observation is a powerful learning mechanism for children (Rogoff et al., 2003). However, little is known about the early development of scientific observation competencies, that is, the ability to systematically use observation as a tool for intentional knowledge seeking. Norris (1984) defines observation competency as the ability to make accurate observations, to report them well and to correctly assess reports of observations. Based on this conceptualization of observation competency as a specific research method, Kohlhauf et al. (2011) developed a competency model, identifying the following dimensions as important for the quality of observation: describing details, questioning, hypothesizing, testing, and interpreting. The authors describe behavior in these dimensions on three ascending levels of incidental observation, unsystematic observation, and systematic observation. In order to validate the model, they analyzed the observation behavior of 110 study participants aged between 4 and 29 years. Kindergarteners were generally on the first level (incidental observation), but even adults did not always reach the third level (systematic observation). The results confirmed a three-dimensional model: describing details, scientific reasoning (questioning, hypothesizing, and testing), and interpreting.

The observation competency model by Kohlhauf et al. (2011) differentiates several important facets in observation: describing, questioning, hypothesizing, testing, and interpreting. As can be seen in **Figure 1**, questioning, hypothesizing, testing, and interpreting are general epistemic activities that are relevant for scientific reasoning processes across domains (Fischer et al., 2014). The cognitive and metacognitive skills needed for these processes are assumed to be domain-general (Piekny et al., 2013b). Only the description of details is specific to observation. Research on observation in the domain of biology has shown that the correct perception and description of relevant details are crucial for making good observations (Eberbach and Crowley, 2009). In this paper, we will treat describing and epistemic activities as two subscales of observation competency.

We ask whether and to what extent kindergarteners' observation of animals, when prompted by an adult, can be described as a scientific reasoning process, characterized by questions/ideas/hypotheses, and the evaluation of observations with respect to these questions or ideas. Further, we ask whether children's domain-independent scientific reasoning competency, as assessed in a knowledge-lean experimentation task, predicts their observation competency in biology, when domain-specific knowledge and other relevant abilities (such as language) are taken into account. Thus, the present study reflects the idea that there is not just one scientific method but several methods with their own structure and difficulties (Lederman et al., 2002). While experimentation and observation are distinct research methods, for both the relevance of both domain-general epistemological understanding and domain-specific knowledge have been shown repeatedly (Chen and Klahr, 1999; Zimmerman, 2007; Eberbach and Crowley, 2009; Kohlhauf et al., 2011).

In general, the quality of an observation has been shown to be strongly influenced by the observer's knowledge in the domain: Observers are influenced by their prior conceptualizations in what they observe or what they think

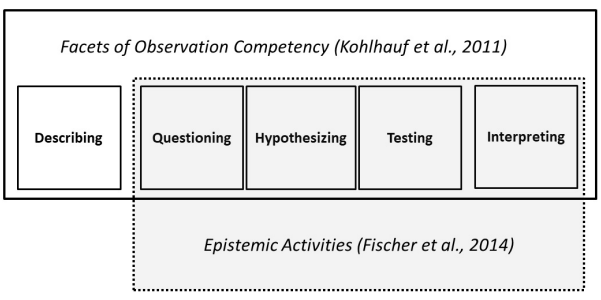


FIGURE 1 | Facets of observation competency and their relation to epistemic activities.

they see (Brewer and Lambert, 2001). Chinn and Malhotra (2002) showed that 75% of students did actually observe incorrectly when the correct observation was not in line with their prior conceptualization. The failure to make correct observations hindered their conceptual change. Yet, there are reasons to assume that observation competency can be a helpful tool in knowledge acquisition and conceptual change (Eberbach and Crowley, 2009). Studies comparing laypersons' and experts' observations show that the two groups use different strategies, with laypersons making several mistakes throughout the whole inquiry process (Eberbach and Crowley, 2009). The experts ask more specific questions and go on questioning and noticing details. Meanwhile, laypersons often ask wrong questions, miss important details, and do not document their observations adequately. Again, this can be seen during the whole inquiry process: domain knowledge is needed to ask the right questions, plan an adequate observation situation, document meaningful details and draw the right conclusions from the data (Alberdi et al., 2000). Kohlhauf et al. (2011) also found that prior knowledge of the object of investigation had a positive impact on the observation competency of their participants, who ranged from kindergarteners to university students. Since domain-specific knowledge has proved to be crucial for children's observation competency (Eberbach and Crowley, 2009; Kohlhauf et al., 2011), we expect children's prior knowledge about the observed objects to have an impact on the performance in the observation situation.

Since the model of observation competency proposed in the present paper places a strong emphasis on epistemic activities involved in the observation process, we further expect domain-general scientific reasoning skills to play a role in the development of observation competency. There is evidence for a development of a domain-general scientific reasoning skill (Osterhaus et al., 2015; Piekny et al., 2013a,b). Thus, we expect to find a correlation between children's grasp of foundational epistemological distinctions and their inquiry skills in an observation situation. While many researchers in science education postulate that domain-specific knowledge is the main motor for the development of scientific skills (Sinatra and Chinn, 2012), we expect both domain-specific knowledge and domain-general understanding of hypotheses and evidence to have an impact on children's performance in a scientific inquiry situation.

Observation competency consists of several facets (compare **Figure 1**). While questioning, hypothesizing, testing, and interpreting are all general epistemic activities and therefore part of all scientific reasoning processes (Fischer et al., 2014), we expect these to show a specific relation with children's domain-general scientific reasoning. The perception and description of relevant details are expected to depend more on prior knowledge in the domain (Eberbach and Crowley, 2009).

There is little investigation of the influence of general cognitive skills on observation competency. However, it has been shown that social interaction is important for developing children's observation skills: in the study by Johnston (2009), children observed and sorted several objects and were interviewed about their procedure. When interacting with peers or adults throughout the task, they showed and reported the use of more sophisticated strategies. Language is important for social interactions and as "intermental (social) activity will promote intramental (individual) intellectual development" (Vygotsky, 1978, p. 86), it can be expected that children with better language skills have experienced more learning situations to improve their reasoning and inquiry skills. Language has not only proved to be an important instrument in the development of false belief understanding (Lohmann and Tomasello, 2003), but also specifically for learning about science (Mercer et al., 2004). Research with children with language impairment suggests that the understanding of causal connectives is crucial for scientific reasoning (Matson and Cline, 2012). A longitudinal study showed that verbal intelligence was positively related to scientific reasoning (Bullock et al., 2009). Children's language abilities have also been found to have an impact on both observation competency (Kohlhauf et al., 2011) and scientific reasoning (Mayer et al., 2014). Therefore, we expect an influence of language on children's reasoning abilities, domain-specific knowledge, and observation competency. As children's executive functions have shown to be related to their scientific reasoning skills (van der Graaf et al., 2016; Osterhaus et al., 2017), we will also measure them in order to be able to control for a potential influence on children's performance in the tasks.

MATERIALS AND METHODS

Sample and Procedure

Eight kindergartens were asked if they wanted to participate in the study, and five actually agreed to do so. Two of these were in an urban environment, three in a rural area. Three of the kindergartens were run by municipal authorities, one by church and one by parent initiative. All kindergartens had basic groups for the children but also group-overarching activities.

We tested 83 children who were in their kindergarten year before starting school. Eight children were excluded from the analyses because their language abilities were so low that the testing could not be run with them as it was with the other children. The cut-off for excluding them from the sample was their performance in the language test. If their results fell into the area of "special educational needs," their performance was not analyzed any further. The age of the final sample of 75 children ranged from 4;9 to 6;7 (years;months); the mean age was 5;6 (65.56 months, SD = 4.67). A total of 38 (51%) of the children were female and 37 (49%) were male.

The testing took place in the kindergartens in a separate room. We tested the children individually in three test blocks, one testing with the observation test, one with the language test and the executive functions test, and one with the scientific reasoning tasks. Each child was usually tested on three different days; only some children were tested twice a day. In these cases, we made sure that they had at least 2 h leisure time in between. The testing either took place at a computer (language test) or was recorded on videotape. If the child did not want to be tested alone, one of the teachers would come along to the testing.

This study was carried out in accordance with the recommendations of the ethics committee of the faculty for

psychology and education at the LMU Munich. As the study was conducted with minors as subjects, all parents or legal guardians gave written informed consent. They had the possibility to withdraw their consent at any time and ask for the deletion of already recorded data. The children themselves also had the possibility to cancel the testing at any time. Parents had the opportunity to ask for their own children's test results.

Instruments

Observation Competency

As mentioned in the introduction, we characterize observation competency as consisting of the two subscales describing and general epistemic activities. The latter subscale in turn consists of the facets questioning, hypothesizing, testing, and interpreting (see **Figure 1**).

For testing children's observation competency, we used the procedure established by Kohlhauf et al. (2011). Here the participants observed a living fish, snail, and mouse. The instructor started the test by introducing a hand puppet and presenting the tools the children could, later on, use for their observations (magnifying glass, ruler, stopwatch, scales, and thermometer). At this point, the animal's cages were still hidden under blankets. After that, the children were shown the first animal, which was always the fish. The puppet closed its eyes and the children were asked to describe the animal to the hand puppet. When the child had finished the description, the puppet opened its eyes again and the experimenter asked the child for a research question ("what do you want to find out about the fish?"). When they had formulated a research question (e.g., "Does the fish have to surface in order to breathe?"), the children had to generate a hypothesis (e.g., "I think the fish must come to get some air"). The children should then observe the animals and try to find answers to their own questions using various aids (e.g., the stopwatch and observe whether or not the fish had to breathe within the selected time). The last step in the observation was to sum up the observation and decide whether to accept or reject the hypothesis that has been set up (e.g., the fish has not surfaced within the set time: does it have to breathe or not?). After observing the fish, the same procedure followed with the snail and lastly the mouse. The whole interaction was videotaped.

The test was designed to find out if the participants are able to describe what they observe, come up with a research question, formulate a hypothesis, do the testing and interpret their observation. Therefore, as little prompting as possible was given by the instructor. If the child got stuck, did not do one of the steps themselves or asked for help, help or prompts were given either by the instructor or the puppet. The need for prompting resulted in scoring less points in the overall score, which will be explained in detail further below.

Table 1 displays two examples of children's actions in the situation. While Child A needed lots of help and prompting, Child B did many steps spontaneously or only needed prompting. As these examples show, Child A did not provide a research question and needed help in order to develop the hypothesis for the given research question. Meanwhile, Child B came up with a usable research question and could form a hypothesis when prompted. In the testing phase, Child A stayed passive and the instructor both gave the idea how to test the question as well as lead the process throughout the observation. Child B, in contrast, came up with the testing idea himself/herself (looking with a magnifying glass) and executed the observation autonomously. Child A did not make a real observation before being prompted to look properly by the instructor. In the beginning of the interpretation, both children could summarize the results when prompted. While Child B also put the results into relation with the hypothesis, Child A failed to do so on his/her own. Neither of the children was able to actively separate between their observation and their interpretation. Examples for this would be any consideration of the limitations of the observation; e.g., stating that the observation would have to be repeated or that the results might be limited to the individual animal instead of being applicable to the whole species, or that the measurement might be imprecise (e.g., when trying to measure the length of the mouse through the glass of the cage).

We first analyzed children's observations according to the procedure of Kohlhauf et al. (2011). In their analysis there were five items (one each for details, questioning, hypothesizing, testing, and interpreting) for each of the three animals, summing up to 15 items in total. This had worked well for their sample that had an age range from kindergarteners to students but proved to be too imprecise for our sample. The children showed floor effects and we were unable to reach satisfactory interrater reliability. We therefore developed a new coding scheme with more items and more gradations within each item. Our final coding scheme for each animal consisted of five facets (describing details, questioning, hypothesizing, testing, and interpreting) with a total of 13 items, which contained up to four gradations. The more autonomous and spontaneous the behavior of the child, the higher the score they were able to achieve. The list of items can be found in **Table 2**, including the scores the examples Child A and Child B from **Table 1** got for their observation of the fish.

The first facet consisted of three items that focused on children's perception of details both during their first description of the animal as well as during the testing phase. One of these three items regarded the number of dimensions (e.g., body parts and overall color) mentioned by the children. The other two items regarded the number of details (e.g., form or color of a specific body part or a description of behavior) that were mentioned by the children. With these two items, we differentiated between "specific" and "unspecific" details to distinguish whether the mentioned details were related to the research question or not. In the examples in **Table 1**, Child A mentioned two dimensions (tail and eyes) and two details for each (tail: silvery with spots, eyes: small and black). Child A did not relate to specific details during the testing phase (only mentioned body parts). Child B described five dimensions (tail fin, green-transparent, eyes, fin on the back, and side fins) and three details related to the tail fin (orange with black dots). As the descriptions of "white and curved" related to the research question on the fishbone, those were counted as two specific details.

The next three facets (questioning, hypothesizing, and testing) consisted of two items each. The items measured both the quality of the performance as well as whether the children performed the

TABLE 1 | Examples of children's behavior in the observation situation and coding.

Phase	Child A		Child B	
	Transcript	**Coding**	**Transcript**	**Coding**
Describing	*Instructor*: Can you describe the fish to Emil (=hand puppet)? *Child A*: A tail that is silvery and with spots on it, it has small eyes, they are black.	Dimensions: 2 Details: 4	*Instructor*: Can you describe the fish to Emil (=hand puppet)? *Child B*: The fish have orange tail fins with black dots, they are green-transparent, they have eyes and a fin on the back and side fins.	Dimensions: 5 Details: 3
Questioning	*Instructor*: Is there something you want to find out about the fish? *Child A*: No. *Instructor*: Okay, I have a question: which fins do they swim with? (Pause)	No question Instructor's question used for investigation	*Instructor*: Is there something you want to find out about the fish? *Child B*: I want to see the very thin stems going up inside or go to the sides, I want to find out how they look (gets up and picks up the magnifying glass) *Instructor*: The fishbone? *Child B*: Yes.	Question when prompted Child's question usable for investigation
Hypothesizing	*Instructor*: What do you think? *Child A*: With the small ones. *Instructor*: Which? Where are they? (Pause) Are they here or here (shows at own body) *Child A*: Here (shows shoulders)	Hypothesis with help	*Instructor*: And what do you think how they look like? *Child B*: Hmmm. . . White, and curved.	Hypothesis when prompted
Testing	*Instructor*: Okay, so let's have a look! (Pause) Which fins do they use? *Child A*: (without really looking) These up here (shows shoulders) *Instructor*: And do other fins move as well? *Child A*: (looks) those down there *Instructor*: Those down at the belly, aha. More? *Child A*: The tail.	Idea: Mostly instructor Execution: Mostly instructor Real observation Specific details: 0	*Instructor*: Okay, so let's have a look. . . you have the magnifier glasses already, but I could also catch the fish in this magnifying glass container. . . *Child B*: Yes! *Instructor*: (catches a fish with the help of *Child B*) *Child B*: (looking at the fish) White and curved! . . . and I now can see a fin on the back that I have not seen before.	Idea: Mostly child Execution: Mostly child Real observation Specific details: 2
Interpreting	*Instructor*: Okay, so what have we seen now? *Child A*: That they move those up here, those down at the belly, and the tail. *Instructor*: And what did you think before? (Pause) Do you remember? *Child A:* (no answer) *Instructor*: So you said they move only those at the shoulders. Is that right? *Child A*: No.	Summary when prompted No relation to hypothesis No separation interpretation/observation	*Instructor*: Okay, so what did you find out? *Child B*: That the fishbone are white and curved. *Instructor*: And what did you think before? *Child B*: The same. But I wanted to be sure.	Summary when prompted Relation to hypothesis on demand No separation interpretation/observation

TABLE 2 | List of items measuring observation competency.

Facets		Items	Example	
			Child A	Child B
Describing details		Dimensions	0.8	1
		Unspecific details	0.63	0.38
		Specific details	0	1
Epistemic activities	Questioning	Research question	0	0.67
		Use of question	0	1
	Hypothesizing	Spontaneous hypothesis	0	0
		Prompted hypothesis	0.67	0.67
	Testing	Activity	0	0.67
		Quality	1	1
	Interpreting	Summary of results	0.5	0.5
		Spontaneous relation to hypothesis	0	0
		Prompted relation to hypothesis	0	0.5
		Differentiation between observation and inferences	0	0

steps spontaneously or if they needed prompting. For this, each item contained 2–4 gradations in order to take the extent of the prompting into account.

Finally, the last facet (interpretation) included four items in which children's summary of the results, their ability to relate them to the hypothesis and the differentiation between observation and inferences were scored. Here again, each item consisted of 2–3 gradations in order to represent the extent to which the children needed prompting. As mentioned above, neither Child A or B differentiated between their observation and their interpretation.

Since the amount of gradations differed between the items, all items were transformed to a value between 0 and 1. In the case of the first facet, in which there were no gradations, the maximal score was transformed to the value of 1 and all other scores were calculated as a percentage. In the other facets the value of 1 represented the highest gradation, e.g., the most autonomous and spontaneous behavior. Therefore, if an item consisted of three gradations, these would be ascribed with the values of 0, 0.5, and 1, respectively.

The 13 items were the same for all three animals, meaning children could reach an overall score between 0 and 39. Child A from our examples (**Table 1**) had an overall score of 11.3, Child B had an overall score of 20.9.

A second rater coded 10% of the data and the Spearman correlations were all above 0.6; for the facets questioning, hypothesizing, testing, and interpreting they were all above 0.9.

As mentioned earlier, we differentiate between the two subscales describing and general epistemic activities (questioning, hypothesizing, testing, and interpreting) (see **Figure 1**). In the analyses, we had a look at the overall scale, the two subscales describing and epistemic activities, and the facets of observation individually (describing, questioning, hypothesizing, testing, and interpreting). The overall scale of observation was reliable ($\alpha = 0.74$). The subscales for details ($\alpha = 0.72$) and the epistemic activities ($\alpha = 0.76$) also showed satisfactory reliability. The values of the facets were only sufficient for questioning ($\alpha = 0.77$) but not for the other facets (hypothesizing: $\alpha = 0.48$, testing: $\alpha = 0.63$, interpreting: $\alpha = 0.40$). We therefore did not conduct any further inference statistics with the facets but will still report the descriptive results.

Scientific Reasoning

We used two tasks to measure children's scientific reasoning abilities: the mouse task by Sodian et al. (1991) and the cake task, which was developed in parallel to the mouse task. Both tasks were administered to the children in form of a story, supported with pictures. Children could point at the pictures to answer but also had to verbally justify their answers. If the justification showed a wrong concept or no justification was given, the answer was coded as wrong. For the mouse task, there were control questions on children's understanding of the task. If the children answered these wrong, their data were coded as missing.

The mouse task: in this task, the children were told the story of two boys who had a mouse in their cellar. The boys had never seen the mouse and therefore did not know if it was big or small. In the first step, they wanted to feed the mouse and had to choose one of two houses (one with a small entrance and one with a big entrance) to put cheese for the mouse in. Hereby the boys wanted to make sure that the mouse could find the food, regardless of its size. In the second step, they wanted to find out if the mouse is big or small and again had to choose one of the two houses to put cheese in. We added a third step, in which we showed the big house, saying the cheese is missing and asked the children if they now knew whether it was a big or a small mouse. With these steps, we assessed our participant's understanding of producing an effect (first step) and of testing a hypothesis using a conclusive test (second and third steps).

The cake task: in this task, a mother baked a cake with two new ingredients and her three children liked the cake a lot. In the first step, the mother wanted to bake the cake again for a birthday party and the children made suggestions what she should do. Hereby the idea was to make sure that the cake tasted the same as the first time (effect production). Child A suggested

to put only one of the ingredients into the new cake, child B suggested to put both ingredients into the cake (right answer), and child C suggested to bake a cake in a square form instead of a round one. In the second step, the mother wanted to find out which of the ingredients is the one to make the cake so tasty because the ingredients were rather expensive and she only wanted to have to buy one. Child A suggested to bake one cake with both ingredients and one cake without both ingredients, Child B suggested to bake one cake with the first and one cake with the second ingredient (right answer), and Child C suggested to bake one round and one square formed cake. In the third step, the family had decided to try out Child A's suggestion and we asked the children if they now found out which ingredient makes the cake tasty. Here again, the second and third steps assessed children's ability to test a hypothesis using a conclusive test.

As we only wanted to analyze children's understanding of testing and not that of producing an effect, we only considered children's answers on the second and third steps of each task, but not their answers on the first steps. Therefore, we had answers to two questions per task, one on the selection of the right answer and one on our additional *post hoc* question. Thus, children could score 0, 1, or 2 points on both scientific reasoning tasks. The frequencies of children's scores are displayed in **Table 3**.

Children's performance on the two tasks was significantly correlated ($\tau = 0.38$, $p < 0.01$), even after language and age had been partialed out ($r = 0.31$, $p < 0.05$). Because of these correlations, we decided to aggregate the two scores to a single scientific reasoning score.

Prior Knowledge Test

We conducted the same test on children's prior knowledge that Kohlhauf et al. (2011) used in their study. The questionnaire consisted of 18 questions about the three animals that were part of the observation situation. The children answered these questions verbally and their answers were written down by the experimenter. Due to floor effects, items that were solved correctly by less than two children had to be deleted. The final scale had 10 items and reached a satisfactory reliability ($\alpha = 0.58$).

Language Abilities

We used the CITO language test (Konak et al., 2005) to measure children's German language abilities. This is a computer-based test to evaluate children's language abilities between age 4;3 and 6;11. The testing took about 25 min. There are four subscales in the test.

In *passive vocabulary*, the children were supposed to click on a picture that displays a word that they were asked to click on. This could either be a noun (e.g., "click on stairs") or a verb (e.g., "click on swimming"). In *cognitive terms*, they also had to click on the right picture, but the content was more complicated. The target could be a color (e.g., "click on white"), the size of an object (e.g., "click on the tallest child"), the number (e.g., "click on the basket with the most apples"), or position of an object (e.g., "click on the house between the trees"). In *phonological awareness,* children heard either two words that sounded very similar or twice the same word. They then had to decide whether it was the same word or two different words. In *text comprehension,* the children heard a short story (4–5 sentences) and afterward were asked multiple-choice questions that tested if they had understood the story correctly and could remember the content. The reliabilities for the subscales were all sufficient to good, being as good as or even better than the ones reported by the authors (see **Table 4**).

Executive Functions

To measure executive functions as an additional control variable, we used the Hearts & Flowers (H&F) task, a computerized test, developed by Diamond (2013). The test was constructed to assess inhibition, set-shifting, and working memory (Davidson et al., 2006; Diamond et al., 2007). The test items were displayed on a computer screen, on which items could be seen on either the left and the right side, and a keyboard with an active key on the left and an active key on the right side. There were three conditions: congruent, incongruent, and mixed. In the congruent condition, children were asked to click the key on the same side as the heart appears. In the incongruent condition, children were asked to click on the opposite side of the flower that appears on the screen. In the mixed condition, either a heart or a flower could appear. If it was a heart, children had to click on the same side; if it was a flower, they had to click on the opposite side. The congruent and incongruent condition each contained 20 trials, the mixed condition contained 33 items.

In the mixed condition, children had to keep both rules in mind (working memory), shift between the rules and inhibit the tendency to press the key on the same side in incongruent items. This condition was therefore the best measure for executive functions and demanding enough to not produce ceiling effects (Zaitchik et al., 2014). Consequently, we used only this scale for our analyses. It reached good reliability ($\alpha = 0.85$).

Statistical Analysis

For the data analysis, we conducted descriptive statistics and calculated correlations and multiple regression analysis using the software IBM SPSS Statistics for Windows, version 24.0. Furthermore, we conducted a mediation analysis using the software PROCESS by Hayes (2012). This program does not

TABLE 3 | Frequencies of scores in the Scientific Reasoning Tasks.

	Cake task	Mouse task
0	31	36
1	16	18
2	19	12

TABLE 4 | Reliabilities of the language test CITO.

Subscale	Test manual	Our sample
Passive vocabulary	0.91	0.89
Cognitive terms	0.88	0.89
Phonological awareness	0.79	0.88
Text comprehension	0.76	0.82
Overall		0.96

only offer to run the Sobel test to determine if the mediation is significant but also gives out bootstrap confidence intervals.

RESULTS

Descriptive Results

The first aim of the analyses was to assess the level of children's observation competency. Additionally to the analysis of the overall observation competency scale, we also had a look at the subscales and facets. For this, we considered describing and the epistemic activities as subscales, as well as all the other facets (questioning, hypothesizing, testing, and interpreting) individually. **Table 5** displays descriptive for all subscales/facets.

In the original study by Kohlhauf (2013), the participants had been sorted into three levels of observation competency. With this categorization, children of that study showed floor effects, as kindergarteners were mostly on the lowest level. Given that, this study could not use the same coding scheme and therefore, the categorization could not be applied in the same way. However, children's performance in the test could still be differentiated in their solving rates of the tasks. For this, we chose to divide the participants in three levels: On level 0, participants scored less than 20% of the points given in the subscale. On level 1, they solved between 20 and 80% of the task. On the highest level (level 2), they solved over 80% of the task correctly. This way, we could better recognize the variability in our sample while examining whether our coding scheme would still elicit floor effects. **Table 6** shows the distribution of the children across the three levels. For the subscales describing and hypothesizing, almost all children were on the medium level. For questioning and testing, the children showed a broader distribution, with testing being the facet with most children in the highest level. Interpreting showed half the children on the lowest and half the children on the medium level.

TABLE 5 | Descriptive for overall observation competency, subscales, and facets.

Variable	*M*	*SD*
Observation competency (sum score)	15.19	4.13
Epistemic activities	0.41	0.11
Describing	0.40	0.15
Questioning	0.43	0.30
Hypothesizing	0.48	0.20
Testing	0.70	0.19
Interpreting	0.20	0.10

TABLE 6 | Descriptive for observation competency levels.

Facet	Level 0 (<20%)	Level 1 (20–80%)	Level 2 (>80%)
Describing	8%	92%	0%
Questioning	24%	59%	17%
Hypothesizing	5%	92%	3%
Testing	3%	68%	29%
Interpreting	49%	51%	0%

Intercorrelations With Cognitive Measures

Table 7 shows the means, standard deviations, and intercorrelations for the overall observation competency measure and the cognitive measures. The expected correlations of observation competency with scientific reasoning, prior knowledge, and language were significant and moderate to strong. Executive functions did not correlate with observation competency. Age did also not have a significant influence on any of the variables except for prior knowledge. Analyses showed that there is no influence by children's gender on the results in any of our measurements.

The intercorrelations of the facets and their correlations with the cognitive measures are displayed in **Table 8**. Both the overall observation competency as well as the two subscales show a moderate positive correlation with language abilities, scientific reasoning, and prior knowledge. There is no significant correlation with executive functions.

Predicting Observation Competency

To further investigate the relations between observation competency and potential influencing factors, we used the significantly correlated variables – scientific reasoning, prior knowledge, and language – as predictors in a multiple regression analysis. All predictors together explained 35% of the variance ($R^2 = 0.35$, $p < 0.001$). The results in **Table 9** show that prior knowledge was the largest influencing factor, followed by domain-general scientific reasoning. Language abilities were not a significant predictor.

Language abilities were also correlated with scientific reasoning and prior knowledge (see **Table 7**), so one assumption is that scientific reasoning and biology understanding mediate the influence of language abilities on observation competency. In order to check for this, we conducted a mediation analysis, using the software PROCESS by Hayes (2012).

We used observation competency as the criterion, language as the independent variable and scientific reasoning and prior knowledge as mediators. There was a significant indirect effect of language on observation competency through both predictors [$b = 0.244$, BCa CI (0.235, 0.365)], as well as a significant indirect effect through only scientific reasoning [$b = 0.153$, BCa CI (0.046, 0.268)], and through only prior knowledge [$b = 0.093$, BCa CI (0.019, 0.180)]. The Sobel test was significant for scientific reasoning ($p = 0.017$), but not for prior knowledge ($p = 0.064$). The results are displayed in **Figure 2**.

Predicting Subscales of Observation Competency

Expecting a specific relation between the general epistemic activities (questioning, hypothesizing, testing, and interpreting) and domain-general scientific reasoning on the one hand and between describing details and prior knowledge, two further regressions were conducted to check these relations. Children's language abilities were kept as a control variable.

Domain-general scientific reasoning, prior knowledge, and language abilities explained 25% of the variance in children's

TABLE 7 | Means, standard deviations, and intercorrelations for observation competency and all predictor variables.

Variable	M	SD	1	2	3	4	5	6
1. Observation competency	15.19	4.13	1					
2. Scientific reasoning	0.37	0.32	0.51**	1				
3. Prior knowledge	0.13	0.15	0.44**	0.42**	1			
4. Language (vocabulary)	0.91	0.09	0.41**	0.47**	0.37**	1		
5. Executive functions	0.70	0.19	0.16	0.50**	0.17	0.45**	1	
6. Age (months)	65.56	4.67	0.20	0.21	0.31**	0.10	−0.06	1

***Significant on 1%-level.*

TABLE 8 | Intercorrelations among subscales and correlations with cognitive measures.

	Observation	Epistemic activities	Describing	Questioning	Hypothesizing	Testing	Interpreting
Observation	1						
Epistemic Activities	0.95**	1					
Describing	0.70**	0.47**	1				
Questioning	0.72**	0.74**	0.34**	1			
Hypothesizing	0.41**	0.44**	0.23*	0.04	1		
Testing	0.52**	0.59**	0.22	0.22	0.05	1	
Interpreting	0.69**	0.69**	0.40**	0.32**	0.15	0.39**	1
Scientific reasoning	0.51**	0.48**	0.38**	0.31**	0.11	0.36**	0.48**
Prior knowledge	0.44**	0.37**	0.45**	0.33**	0.10	0.27*	0.30**
Language	0.41**	0.36**	0.28*	0.15	0.32**	0.18	0.36**
Executive functions	0.16	0.17	0.05	0.16	0.12	0.01	0.18

**Significant on 5%-level; **Significant on 1%-level.*

describing (R^2 = 0.25, $p < 0.001$) and 28% of the variance in children's performance throughout the epistemic activities (R^2 = 0.28, $p < 0.001$). For both scales, language abilities were not a significant predictor (see **Tables 10**, **11**). While prior knowledge was the only significant predictor for describing, scientific reasoning was the only significant predictor for epistemic activities.

TABLE 9 | Regression analysis summary for scientific reasoning, prior knowledge, and language predicting observation competency.

Variable	*B*	*SE B*	β	*t*	*p*
(Constant)	5.78	4.41		1.31	0.20
Scientific reasoning	4.13	1.44	0.33	2.88	0.00
Prior knowledge	6.98	3.02	0.25	2.31	0.02
Language	7.60	5.12	0.17	1.48	0.14

DISCUSSION

The present study is the first systematic investigation of scientific observation competency in young children. Observation competency was defined as comprising the ability to describe features of target animals, as well as to generate questions and hypotheses with regard to the target animals, to test these hypotheses, and to interpret the findings with respect to the question or hypothesis (epistemic activities). Our first aim was to describe the scope and limits of observation competency in kindergarteners, while the second aim was to relate individual differences in children's observation competencies to general cognitive abilities (e.g., language and executive functions), domain-general scientific reasoning skills and domain-specific knowledge.

The descriptive data indicated that there were no floor effects in kindergarteners for most facets of observation competency. One exception was the ability to generate interpretations for their observations: on this facet, about half of the sample did not respond even when prompted. Most children showed evidence – at least when prompted – for some epistemic activities. It is possible (and remains to be explored further) that the differentiation of data and interpretation is harder to grasp for young children than the basic idea of testing hypotheses through specific observations.

Given that there was both a sufficient level of performance and individual variability with respect to observation competency in the present sample, it was possible to investigate the predictors of observation competency in kindergarteners.

Our hypothesis was that both children's domain knowledge and their domain-general scientific reasoning ability would have an effect on their observation competency even if more general cognitive abilities were controlled for. These hypotheses were corroborated by the data: scientific reasoning proved to be a significant predictor for children's observation competency alongside with children's prior knowledge about animals. The results thus indicated that not only domain-specific competencies are important for scientific observation, but also domain-general reasoning abilities.

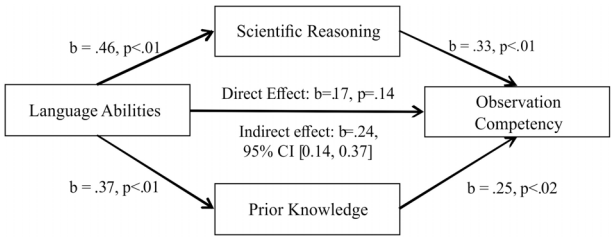


FIGURE 2 | Mediation analysis of the indirect effect of language on observation competency through scientific reasoning and prior knowledge.

TABLE 10 | Regression analysis summary for scientific reasoning, prior knowledge, and language predicting general epistemic activities in observation competency.

Variable	*B*	*SE B*	β	*t*	*p*
(Constant)	0.18	0.13		1.44	0.16
Scientific reasoning	0.12	0.04	0.33	2.78	0.01
Prior knowledge	0.14	0.09	0.19	1.64	0.11
Language	0.18	0.15	0.14	1.20	0.23

TABLE 11 | Regression analysis summary for scientific reasoning, prior knowledge, and language predicting describing in observation competency.

Variable	*B*	*SE B*	β	*t*	*p*
(Constant)	0.24	0.17		1.45	0.15
Scientific reasoning	0.10	0.06	0.22	1.77	0.08
Prior knowledge	0.34	0.12	0.34	2.90	0.01
Language	0.09	0.20	0.05	0.44	0.67

With respect to the effects of domain knowledge, the present findings once again demonstrate that "if it is true that thinking and reasoning are processes, so too it is true that content knowledge is the fodder for these processes" (Sinatra and Chinn, 2012, p. 258). Content knowledge was assessed as simple factual knowledge about animals in the present study. Therefore, a linear positive relation between the amount of content knowledge and the ability to describe concrete observations in animals was expected. Our findings are not inconsistent with more complex models of the interaction between children's content knowledge and reasoning or observation skills which emphasize that prior domain-specific beliefs can be an impediment to the coordination of theory and evidence (Koerber et al., 2005; Croker and Buchanan, 2011). Research has also shown that evidence contradicting children's prior beliefs can also lead to further and deeper inquiry (Legare et al., 2010). Further research is necessary to determine the effects of children's conceptual understanding in the domain of biology (e.g., childhood animism and concept of living things) on young children's observation competencies. Children's naïve concepts of living things (Gelman, 2009) and their tendency to focus on goal-direction (Evans, 2008) would be factors that could hamper children's observation competency, while more sophisticated knowledge about the domain may lead to better, unbiased reasoning (Geary, 2008).

The finding that the ability to distinguish between a conclusive and an inconclusive test for a simple hypothesis in an everyday domain predicted children's observation competency in biology was predicted on the grounds that the differentiation of hypotheses from evidence is assumed to be fundamental for scientific reasoning in general, not just for experimentation skills. It should be noted that the observation competency assessment did not include the notion of a conclusive test and was not similar in terms of task demands to the scientific reasoning task. Thus, it appears that hypothesis – evidence differentiation is a metaconceptual distinction that underlies a wide range of scientific reasoning abilities, and that is domain-independent. The finding that the score attained for the general epistemic activities, not for the description of specific details, was related to children's domain-general scientific reasoning also supports this interpretation. The perception and description of relevant details, on the other hand, was more closely related to children's domain-specific knowledge. Eberbach and Crowley (2009) argue that laypersons with scarce domain-specific knowledge often miss the meaningful details or concentrate on irrelevant properties of the observed object.

Furthermore, we found a correlation between scientific reasoning and executive functions. This result is consistent with a growing body of findings indicating an association of scientific reasoning and executive function measures in different age groups (Mayer et al., 2014; Osterhaus et al., 2017). However, executive functions showed no relation to observation competency or any of its subscales, at least with the executive functions task used in the study. This could suggest that it is elicited reasoning, rather than spontaneous response tasks that show higher executive demands, as children's spontaneous reactions were recorded before prompts were given. Observation competency was linked to scientific reasoning independently of executive functions, thus again supporting the idea that the metaconceptual understanding of the hypothesis evidence relation is foundational for a wide range of scientific activities. Still, other executive functions tasks with a higher emphasis on working memory or planning abilities might show a direct relation to observation. Further research is needed to better understand this relation.

We also assumed that children's language abilities have an influence on all of the other measures. While we did find correlations between language and reasoning abilities, domain knowledge and observation competency, language ability was not a significant predictor for observation competency when we did a regression analysis with all three predictors. Our mediation analysis showed that the influence of language on children's performance in the observation task was mediated by both domain-general and domain-specific science skills. This finding appears to be consistent with the interpretation that children's language abilities influence both general reasoning abilities and knowledge, which both contribute to children's abilities in a concrete observation situation. This also means that the impact scientific reasoning and prior knowledge have on the observation competency is more than just the shared influence of language: they both had a specific, independent effect on children's performance in the observation task. Sociocultural

theories postulate that intellectual competencies are a cultural product and are therefore derived through social interaction (Vygotsky, 1978; Tomasello, 2013). Our results could fortify these theories – language was an influencing factor on all our measurements. However, language seems to have a more direct influence on knowledge and reasoning, while these then shape the behavior in the scientific inquiry situation. Of course, it is also possible that the effect of language we found was a testing effect – as all our instruments were, of course, language-based, we cannot refute this alternative explanation. In this case, however, the influence of both knowledge and reasoning is more than an effect of the verbal testing method because their relation to observation competency stays significant when controlling for language.

In sum, the present study has shown that both domain-specific knowledge and domain-general scientific reasoning abilities contribute to children's observation competency in the domain of biology. This is notable since metaconceptual foundations of scientific reasoning only begin to develop in this age group. Further research is needed to determine the interrelations of these core components of scientific activities over a wider age range. The study focused on scientific observation. Many aspects of observation competency as defined in this study are general epistemic activities, such as hypothesizing and interpreting observations. Further research is needed to determine the generalizability of the present findings to other scientific methods.

DATA AVAILABILITY STATEMENT

The dataset for this study can be found on OSF: https://osf.io/tszay/ (Klemm et al., 2020). The raw data (video recordings) cannot be made publicly available due to the protection of privacy of the study participants. Further materials have also been published in the appendices of the dissertation of Janina Klemm, "Biological Observation Competency in Preschool: the Relation to Scientific Reasoning and Opportunities for Intervention" (Klemm, 2017), accessible on https://edoc.ub.uni-muenchen.de/. The instruments used in this study can be made available to interested researchers upon request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics committee of the faculty for psychology and education at the LMU, Munich. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

JK, BN, and BS contributed to the conception and design of the study. JK organized the database, performed the statistical analysis, and wrote the first draft of the manuscript. All authors wrote sections of the manuscript, contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was kindly supported by the Elite Network of Bavaria (grant number K-GS-2012-209).

ACKNOWLEDGMENTS

The authors would like to thank Özgün Köksal-Tuncer for her contributions to the development of the scientific reasoning task, Lucia Kohlhauf for her help and feedback in all stages of the study, and the kindergartens and families who participated in the research. The data and results presented in this study were part of a dissertation project and therefore are also presented in the final thesis, "Biological Observation Competency in Preschool: the Relation to Scientific Reasoning and Opportunities for Intervention" by Janina Klemm (Klemm, 2017), accessible on https://edoc.ub.uni-muenchen.de/25414/. The dissertation contains partly different analyses. While parts of the results presented in this paper have been presented at conferences, the present analyses significantly exceed the previous ones presented.

REFERENCES

Alberdi, E., Sleeman, D. H., and Korpi, M. (2000). Accommodating surprise in taxonomic tasks: the role of expertise. *Cogn. Sci.* 24, 53–91. doi: 10.1207/s15516709cog2401_2

Brewer, W. F., and Lambert, B. L. (2001). The theory-ladenness of observation and the theory-ladenness of the rest of the scientific process. *Philos. Sci.* 68(Suppl. 3), S176–S186.

Bullock, M., Sodian, B., and Koerber, S. (2009). "Doing experiments and understanding science: development of scientific reasoning from childhood to adulthood," in *Human Development from Early Childhood to Early Adulthood. Findings from the Munich Longitudinal Study*, eds W. Schneider and M. Bullock (Mahwah, NJ: Erlbaum), 173–197.

Chen, Z., and Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Dev.* 70, 1098–1120. doi: 10.1007/978-1-4020-5814-1_1

Chinn, C. A., and Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: a theoretical framework for evaluating inquiry tasks. *Sci. Educ.* 86, 175–218.

Croker, S., and Buchanan, H. (2011). Scientific reasoning in a real-world context: the effect of prior belief and outcome on children's hypothesis-testing strategies. *Br. J. Dev. Psychol.* 29, 409–424. doi: 10.1348/026151010X496906

Davidson, M. C., Amso, D., Anderson, L. C., and Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia* 44, 2037–2078.

Diamond, A. (2013). Executive functions. *Annu. Rev. Psychol.* 64, 135–168. doi: 10.1146/annurev-psych-113011-143750

Diamond, A., Barnett, W. S., Thomas, J., and Munro, S. (2007). Preschool program improves cognitive control. *Science* 318, 1387–1388. doi: 10.1126/science.1151148

Eberbach, C., and Crowley, K. (2009). From everyday to scientific observation: how children learn to observe the biologist's world. *Rev. Educ. Res.* 79, 39–68. doi: 10.3102/0034654308325899

Evans, E. M. (2008). "Conceptual change and evolutionary biology: a developmental analysis," in *International Handbook of Research on Conceptual Change*, ed. S. Vosniadou (New York, NY: Routledge), 263–294.

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., et al. (2014). Scientific reasoning and argumentation: advancing an interdisciplinary research agenda in education. *Frontline Learn. Res.* 2:28–45. doi: 10.14786/flr.v2i2.96

Geary, D. C. (2008). An evolutionarily informed education science. *Educ. Psychol.* 43, 179–195. doi: 10.1080/00461520802392133

Gelman, S. A. (2009). Learning from others: children's construction of concepts. *Annu. Rev. Psychol.* 60, 115–140. doi: 10.1146/annurev.psych.59.103006.093659

Gopnik, A., and Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychol. Bull.* 138, 1085–1108. doi: 10.1037/a0028044

Hayes, A. F. (2012). PROCESS: A Versatile Computational Tool for Observed Variable Mediation, Moderation, and Conditional Process Modeling. Availbale online at: http://www.afhayes.com/public/process2012.pdf (accessed May 5, 2020).

Johnston, J. S. (2009). What does the skill of observation look like in young children? *Int. J. Sci. Educ.* 31, 2511–2525. doi: 10.1080/09500690802644637

Klemm, J. (2017). *Biological Observation Competency in Preschool – the Relation to Scientific Reasoning and Opportunities for Intervention*. Ph.D. dissertation, Ludwig-Maximilians-Universität München, Munich.

Klemm, J., Flores, P. S., Sodian, B., and Neuhaus, B. J. (2020). *Scientific Reasoning in Biology*. Available online at: osf.io/tszay (accessed January 16, 2020).

Koerber, S., Sodian, B., Thoermer, C., and Nett, U. (2005). Scientific reasoning in young children: preschoolers' ability to evaluate covariation evidence. *Swiss J. Psychol.* 64, 141–152. doi: 10.1024/1421-0185.64.3.141

Kohlhauf, L. (2013). *Spielend Biologisch Beobachten: Messung und Förderung Biologischer Beobachtungskompetenz im Vorschulalter*. Berlin: Logos-Verlag.

Kohlhauf, L., Rutke, U., and Neuhaus, B. (2011). Influence of previous knowledge, language skills and domain-specific interest on observation competency. *J. Sci. Educ. Technol.* 20, 667–678. doi: 10.1007/s10956-011-9322-3

Konak, Ö, Duindam, T., and Kamphuis, F. (2005). *Cito-Sprachtest. Wissenschaftlicher Bericht. Retrieved January, 28, 2006*. Butzbach: Cito Deutschland.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychol. Rev.* 96, 674–689. doi: 10.1037/0033-295X.96.4.674

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., and Schwartz, R. S. (2002). Views of nature of science questionnaire: toward valid and meaningful assessment of learners' conceptions of nature of science. *J. Res. Sci. Teach.* 39, 497–521. doi: 10.1002/tea.10034

Legare, C. H., Gelman, S. A., and Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Dev.* 81, 929–944.

Leuchter, M., Saalbach, H., and Hardy, I. (2014). Designing science learning in the first years of schooling. An intervention study with sequenced learning material on the topic of "floating and sinking.". *Int. J. Sci. Educ.* 36, 1751–1771. doi: 10.1080/09500693.2013.878482

Lohmann, H., and Tomasello, M. (2003). The role of language in the development of false belief understanding: a training study. *Child Dev.* 74, 1130–1144.

Matson, G., and Cline, T. (2012). The impact of specific language impairment on performance in science and suggested implications for pedagogy. *Child Lang. Teach. Ther.* 28, 25–37.

Mayer, D., Sodian, B., Koerber, S., and Schwippert, K. (2014). Scientific reasoning in elementary school children: assessment and relations with cognitive abilities. *Learn. Instr.* 29, 43–55. doi: 10.1016/j.learninstruc.2013.07.005

Mercer, N., Dawes, L., Wegerif, R., and Sams, C. (2004). Reasoning as a scientist: ways of helping children to use language to learn science. *Br. Educ. Res. J.* 30, 359–377.

Norris, S. P. (1984). Defining observational competence. *Sci. Educ.* 68, 129–142. doi: 10.1002/sce.3730680206

Oguz, A., and Yurumezoglu, K. (2007). "The primacy of observation in inquiry-based science teaching," in *Paper presented at the International Workshop: Science Education in School*, Bucharest.

Osterhaus, C., Koerber, S., and Sodian, B. (2015). Children's understanding of experimental contrast and experimental control: an inventory for primary school. *Frontline Learn. Res.* 3:56–94. doi: 10.14786/flr.v3i4.220

Osterhaus, C., Koerber, S., and Sodian, B. (2017). Scientific thinking in elementary school: children's social cognition and their epistemological understanding promote experimentation skills. *Dev. Psychol.* 53, 450–462.

Piekny, J., Grube, D., and Maehler, C. (2013a). The development of experimentation and evidence evaluation skills at preschool age. *Int. J. Sci. Educ.* 2, 1–21. doi: 10.1080/09500693.2013.776192

Piekny, J., Grube, D., and Maehler, C. (2013b). The relation between preschool children's false-belief understanding and domain-general experimentation skills. *Metacogn. Learn.* 8, 103–119. doi: 10.1007/s11409-013-9097-4

Rogoff, B., Paradise, R., Arauz, R. M., Correa-Chávez, M., and Angelillo, C. (2003). Firsthand learning through intent participation. *Annu. Rev. Psychol.* 54, 175–203. doi: 10.1146/annurev.psych.54.101601.145118

Sinatra, G. M., and Chinn, C. A. (2012). "Thinking and reasoning in science: promoting epistemic conceptual change," in *APA Educational Psychology Handbook, Vol 3: Application to Learning and Teaching*, eds K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, and H. L. Swanson (Washington, DC: American Psychological Association), 257–282. doi: 10.1037/13275-011

Sodian, B., Zaitchik, D., and Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Dev.* 62, 753–766. doi: 10.1111/j.1467-8624.1991.tb01567.x

Tomasello, M. (2013). *Die Kulturelle Entwicklung des Menschlichen DENKENS: zur Evolution der Kognition*. Berlin: Suhrkamp Verlag.

van der Graaf, J., Segers, E., and Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: dynamic assessment of the control of variables strategy. *Instr. Sci.* 43, 381–400. doi: 10.1007/s11251-015-9344-y

van der Graaf, J., Segers, E., and Verhoeven, L. (2016). Scientific reasoning in kindergarten: cognitive factors in experimentation and evidence evaluation. *Learn. Individ. Diff.* 49, 190–200. doi: 10.1016/j.lindif.2016.06.006

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Mental Process*. Cambridge, MA: Harvard University Press.

Zaitchik, D., Iqbal, Y., and Carey, S. (2014). The effect of executive function on biological reasoning in young children: an individual differences study. *Child Dev.* 85, 160–175. doi: 10.1111/cdev.12145

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Dev. Rev.* 27, 172–223.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Klemm, Flores, Sodian and Neuhaus. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 05 June 2020
doi: 10.3389/fpsyg.2020.01123

Children's Spontaneous Gestures Reflect Verbal Understanding of the Day/Night Cycle

Caroline M. Gaudreau[1], Florencia K. Anggoro[2]* and Benjamin D. Jee[3]*

[1] College of Education and Human Development, University of Delaware, Newark, DE, United States, [2] Department of Psychology, College of the Holy Cross, Worcester, MA, United States, [3] Department of Psychology, Worcester State University, Worcester, MA, United States

Understanding the day/night cycle requires integrating observations of the sky (an *Earth-based* perspective) with scientific models of the solar system (a *space-based* perspective). Yet children often fail to make the right connections and resort to non-scientific intuitions – for example, the Sun moving up and down – to explain what they observe. The present research explored whether children's *gestures* indicate their conceptual integration of Earth- and space-based perspectives. We coded the spontaneous gestures of 85 third-grade children in U.S. public schools (M_{age} = 8.87 years) as they verbally explained the overall cause of the day/night cycle, the cause of sunrise, and the cause of sunset after receiving science instruction as part of a prior study. We focused on two kinds of gestures: those reflecting the Sun's motion across the sky and those reflecting the Earth's axial rotation. We found that participants were more likely to produce Earth rotation gestures for a topic they explained *more* accurately (the overall cause of the day/night cycle), whereas Sun motion gestures were more common for topics they explained *less* accurately (the causes of sunrise and sunset). Further, participants who produced rotation gestures tended to provide more accurate verbal explanations of the overall cause. We discuss how gestures could be used to measure – and possibly improve – children's conceptual understanding and why sunrise and sunset may be particularly difficult topics to learn.

Keywords: gesture, astronomy, mental models, day/night cycle, embodiment

OPEN ACCESS

Edited by:
Stella Vosniadou, Flinders University, Australia

Reviewed by:
Julia Plummer, The Pennsylvania State University, United States
Eva E. Chen, The Hong Kong University of Science and Technology, Hong Kong

****Correspondence:***
Caroline M. Gaudreau
cmorano@udel.edu
Florencia K. Anggoro
fanggoro@holycross.edu

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *15 January 2020*
Accepted: *01 May 2020*
Published: *05 June 2020*

Citation:
Gaudreau CM, Anggoro FK and Jee BD (2020) Children's Spontaneous Gestures Reflect Verbal Understanding of the Day/Night Cycle. Front. Psychol. 11:1123. doi: 10.3389/fpsyg.2020.01123

INTRODUCTION

Promoting student participation and performance in STEM (science, technology, engineering, and mathematics) is an educational priority in the United States. To help more students succeed, we must better understand how students think and learn in these disciplines. In this paper, we explore how children spontaneously gesture when expressing ideas about an intensely spatial science topic, the day/night cycle. These non-verbal behaviors may reflect children's thinking about spatial systems and, if better understood, provide a means for influencing conceptual understanding.

There is mounting evidence that spatial thinking – including mental rotation, mental transformation, and perspective taking – contributes to STEM learning outcomes (Wai et al., 2009; Lee and Bednarz, 2012; Uttal and Cohen, 2012; Heywood et al., 2013; Newcombe and Shipley, 2015; Stieff et al., 2016; Carr et al., 2017). Consider space science. To grasp fundamental ideas

such as the day/night cycle, a student must connect observations from the Earth's surface (e.g., the Sun appearing to rise in the sky) with large-scale events in our solar system (e.g., a location on the Earth becoming exposed to sunlight due to the Earth's rotation). Mentally integrating Earth- and space-based perspectives of the solar system represents a difficult spatial thinking challenge (Heywood et al., 2013; Plummer, 2014, 2016). Besides the difficulty of mapping Earth- and space-based perspectives of the solar system, children's own intuitions about the world (e.g., witnessing the Sun's apparent motion) can impede their understanding (Shtulman, 2017). As children attempt to map the relations between Earth- and space-based perspectives, they may rely on inherently spatial modes of expression, such as *gesture*. Gestures often add information that is missing from verbal explanations (Goldin-Meadow, 2005, 2015; Alibali and Kita, 2010; Alibali et al., 2011, 2013; Özçalışkan et al., 2014; Waters and Beck, 2015). Highly visuospatial concepts in STEM may be more easily expressed through gesture than speech (Iverson and Goldin-Meadow, 2005; Atit et al., 2015; Stieff et al., 2016). Gestures can reduce demands on memory systems (Iverson and Goldin-Meadow, 2005) and can increase focus on a topic (Goldin-Meadow, 2015). In fact, certain concepts may depend on encoding through bodily motion and corresponding sensorimotor brain regions (Matthews-Saugstad et al., 2016; Alibali and Nathan, 2018). During learning about the day/night cycle, children can use gestures to model real or apparent motion from different frames of reference (Plummer et al., 2016).

The purpose of the present study was to explore gestures that children spontaneously produced while explaining key events in the day/night cycle. Our sample included dozens of third-grade children in U.S. public schools who were interviewed before and after completing a series of lessons about the day/night cycle as part of an earlier study (Jee and Anggoro, 2019). While the study involved three different instructional conditions, all conditions covered the same concepts and involved instruction with a 3D model of the Earth–Sun system (see Jee and Anggoro, 2019, for further detail).

The current analyses focused on participants' responses to questions about the cause of (1) the overall change from day to night, (2) sunrise, and (3) sunset. If a participant grasped the scientific explanations from the lessons, then they should provide the same causal explanation for all three topics – namely, the Earth's eastward rotation as seen from a space-based perspective. However, if a participant is focused on the Sun's apparent motion from an Earth-based perspective, or confused about whether the Sun moves, they may provide inadequate or incomplete verbal explanations. Such confusions and omissions may be more frequent for sunrise and sunset, which invoke an Earth-based frame of reference and are labeled in terms of the Sun motion, than for the overall day/night cycle.

We coded participants' non-verbal behavior as they responded to relevant interview questions, looking specifically for the occurrence of two kinds of gestures: (1) *Earth rotation gestures* that represent the rotating motion of the Earth from a large-scale, space-based perspective and (2) *Sun motion gestures* that indicate movement of the Sun across the sky. We considered two main ways in which gestures could relate to verbal understanding:

1. Gesture as a *mirror* that reflects *existing*, verbalized knowledge (e.g., Alibali, 2005; Plummer et al., 2016).
2. Gesture as a *window* into ideas that are not yet (or cannot be) expressed in speech. In this sense, gestures indicate burgeoning conceptual change, predicting *future* breakthroughs in verbalized knowledge (e.g., Church and Goldin-Meadow, 1986; Goldin-Meadow, 2003).

If gesture acts like a *mirror*, then Earth rotation gestures should be especially frequent when participants verbalize high levels of causal understanding, and Sun motion gestures should be more frequent when verbalized understanding is low. A finer-grained prediction is that children who make Earth rotation gestures should explain the day/night cycle more accurately than should children who do not, both at pretest and at posttest. If, however, gestures are a *window* into emerging knowledge, then Earth rotation gestures at pretest should predict verbalized understanding following instruction, at the posttest.

MATERIALS AND METHODS

Participants

Participants included 85 third-grade U.S. public school students (M_{age} = 8.87, SD = 0.50, 57% female) who completed a series of lessons about the day/night cycle as part of a larger study (Jee and Anggoro, 2019) between October 2015 and December 2016. The demographic distribution of this sample reflected that of the school district from which the sample was acquired, 14.9% were African American, 7.5% Asian, 40.8% Hispanic, 0.2% Native American, 32.5% White, and 4.1% multiracial, non-Hispanic.

Interviews and Coding Procedure

All participants completed pretest and posttest interviews about the day/night cycle. Each interview lasted about 15–20 min per child and was videotaped by a trained research assistant. Interview questions were administered verbally by the research assistant. The questions asked participants to explain the cause of (1) the overall change from daytime to nighttime, (2) sunrise, and (3) sunset (see Jee and Anggoro, 2019, for further details). **Table 1** provides the relevant interview questions along with the two knowledge components that were used to score participants' verbal responses for each topic. Children scored 1 point for each knowledge component they correctly verbalized. Intercoder reliability for verbal knowledge scoring was established through independent coding trials, followed by reliability analyses, discussion, and refinement of the coding criteria. All coders obtained reliability of 0.80 or higher (Krippendorff's α) with the other coders on two consecutive rounds of four to six interviews. Reliability ranged from 0.82 to 0.96 across all knowledge components (Jee and Anggoro, 2019).

Children's gestures during the interviews were also coded. Rules for coding gestures were made stringent: (a) a complete gesture necessitated a clear break from any fidgeting of the hand or fingers before or after the gesture, and (b) any gestures made toward technology or items in the classroom were not included, on the basis that not all children had the same resources

TABLE 1 | Interview questions and knowledge components for scoring responses.

	Topic		
	Overall cause	**Sunrise**	**Sunset**
Question(s)	What causes the change from daytime to nighttime?	Every day in [a large city in Northeastern United States], an event happens that we call "sunrise." It looks like this [show video of sunrise]. Have you seen sunrise before? What is sunrise? Why does sunrise happen? Why does sunrise happen in the east?	Every day in [a large city in Northeastern United States], an event happens that we call "sunset." It looks like this [show video of sunset]. Have you seen sunset before? What is sunset? Why does sunset happen? What makes sunset happen in the west?
Knowledge components	The Earth spins/turns/rotates Rotation causes places to face the Sun at different times.	The Earth rotates east/a place rotates toward the Sun. The Earth rotates until a place begins to face the Sun.	The Earth rotates east/a place rotates away from the Sun. The Earth rotates until a place begins to face away from the Sun.

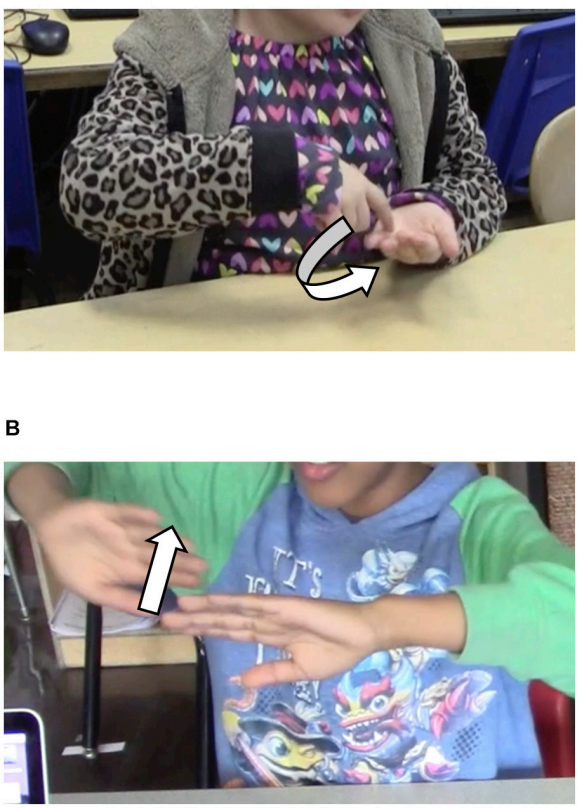


FIGURE 1 | Examples of **(A)** Earth rotation and **(B)** Sun motion gestures. Arrow indicate movement of finger or hand.

available during the interviews. Gestures were categorized as one of two types (see **Figure 1**), based largely on work by Goldin-Meadow (2005) and Plummer (2014). *Earth rotation* gestures involved children using the finger, hand, or arm to represent the rotating motion of the Earth. For the purpose of this study, only instances where children explicitly mentioned "the Earth" in their verbal responses were considered when coding rotation gestures, to avoid ambiguity concerning what children's gestures were intended to represent. In *Sun movement* gestures, children used the finger, hand, or arm to indicate leftward, rightward, upward, or downward movement of celestial objects. In nearly all instances observed, these gestures referenced the Sun motion (i.e., children mentioned "the Sun" in their verbal response). Coders made notes during coding, so only movement gestures that included an explicit verbal reference to the Sun were analyzed for the current study. The coders established interrater reliability (Krippendorff's $\alpha = 0.83$) using 10% of the 170 pretest and posttest interviews.

RESULTS

Verbal Explanations

Participants' knowledge of the three topics was scored in terms of the number of components they included in their verbal explanations (see **Table 2**). We conducted a 2 (Session: pre, post) × 3 (Topic: overall cause, sunrise, and sunset) repeated-measures ANOVA to compare children's knowledge of the overall cause of the day/night cycle, sunrise, and sunset at pretest and posttest. Assumptions for Mauchly's *W* were not met, so results from the Huynh–Feldt model are reported, as suggested by Howell (2002) and Field (2013). Results revealed

TABLE 2 | Descriptive statistics for children's knowledge and gestures by topic.

Variable	**Topic**		
	Overall cause	**Sunrise**	**Sunset**
Knowledge components			
Pretest	0.72 (0.77)	0.05 (0.26)	0.05 (0.21)
Posttest	1.35 (0.70)	0.41 (0.64)	0.39 (0.62)
Earth rotation gesture			
Pretest	32 (37.6%)	15 (17.6%)	17 (20.0%)
Posttest	32 (37.6%)	18 (21.2%)	14 (16.5%)
Sun motion gesture			
Pretest	19 (22.4%)	37 (43.5%)	41 (48.2%)
Posttest	6 (7.1%)	29 (34.1%)	23 (27.1%)

N = 85. Number of children (and percentage) who made a gesture reported for each type of gesture.

main effects of Session, $F(1, 84) = 25.48$, $p < 0.001$ (participants had higher knowledge at posttest than pretest), and Topic, $F(1.45, 121.48) = 177.71$, $p < 0.001$ [participants had higher knowledge for the overall cause than for sunrise ($p < 0.001$) and sunset ($p < 0.001$) but did not differ in their knowledge of sunrise and sunset ($p = 0.697$)]. There was also a significant interaction between Session and Topic, $F(1.58, 132.73) = 5.24$, $p = 0.011$ – although knowledge of each topic increased significantly, $Fs > 13.0$, $ps < 0.001$, the increase was greater for the overall cause of the day/night cycle than for sunrise or sunset.

Relations Between Verbal Explanations and Gestures

Gestures were coded categorically; children received a 1 if they made a given type of gesture when verbally explaining a specific topic and a 0 if they did not make the gesture. **Table 2** shows the number of children who made a given type of gesture for each topic at pretest and posttest. Earth rotation gestures appear to be more common during explanations of the overall cause than for sunrise and sunset. Sun motion gestures, however, were more likely during explanations of sunrise and sunset. This pattern for Sun motion gestures appears to persist from pretest to posttest despite an overall decrease in the frequency of Sun motion gestures.

We planned to test how gestures related to verbally expressed knowledge by comparing the average knowledge of participants who did vs. did not produce a given type of gesture for each topic. However, gesture frequency was less than 25% in some cases (e.g., only 17.6% of participants made a rotation gesture for the sunrise topic at pretest). To avoid comparing wildly uneven group sizes and the consequent loss of statistical power (e.g., Rusticus and Lovato, 2014), we divided participants into Earth rotation gesturers and non-gesturers on the basis of their gestures for the overall cause topic – the topic for which rotation gestures were most frequent[1]. We then tested whether participants who made an Earth rotation gesture for the overall cause topic differed in their verbally expressed knowledge for each of the three topics. We applied a similar rationale to divide participants into Sun motion gesturers and non-gesturers. Specifically, we tested whether participants who made a Sun movement gesture while describing sunrise or sunset – topics for which Sun motion gestures were most frequent – differed in their verbally expressed knowledge for each of the three topics. **Table 3** shows the mean verbal knowledge scores on each topic for the Earth rotation and Sun motion gesture groupings.

Earth Rotation Gestures

At pretest, participants who made an Earth rotation gesture for the overall cause topic tended to have higher knowledge of the overall cause than did participants who did not make the rotation gesture, $t(83) = 2.10$, $p = 0.039$. However, these rotation gesturers had about the same level of knowledge of sunrise as participants who did not make the rotation gesture, $t(83) = -1.28$, $p = 0.205$. Verbal knowledge of sunset was very low overall at pretest. Rotation gesturers expressed no knowledge of sunset, whereas non-rotation gesturers had slightly higher verbal knowledge, $t(52) = 1.60$, $p = 0.044$ (adjusting degrees of freedom in light of unequal variances in Levene's test, $F = 12.10$, $p = 0.001$).

At posttest, participants who made an Earth rotation gesture for the overall cause topic tended to have higher knowledge of the overall cause than did those who did not make the rotation gesture, $t(82.52) = -2.04$, $p = 0.044$ (adjusting degrees of freedom in light of unequal variances in Levene's test, $F = 8.04$, $p = 0.006$). These rotation gesturers had about the same level of knowledge of sunrise as participants who did not make the rotation gesture, $t(83) = -0.91$, $p = 0.528$, and likewise for sunset, $t(83) = 0.513$, $p = 0.610$.

Sun Motion Gestures

At pretest, participants who made the Sun motion gesture when explaining sunrise had about the same knowledge level of the overall cause topic as participants who did not make the Sun motion gesture, $t(83) = -0.98$, $p = 0.328$. These Sun motion gesturers also expressed about the same level of knowledge of sunrise as the non-gesturers, $t(83) = -0.614$, $p = 0.541$, and likewise for sunset, $t(83) = 0.759$, $p = 0.450$.

At posttest, participants who made a Sun motion gesture for the sunrise topic had about the same knowledge level of the overall cause as participants who did not make the Sun motion gesture, $t(83) = 0.726$, $p = 0.470$. These Sun motion gesturers also

TABLE 3 | Mean verbal knowledge of each topic for gesture and non-gesture groups at pretest and posttest.

Gesture grouping	Topic		
	Overall cause	Sunrise	Sunset
Earth rotation gesture for overall cause		**Pretest**	
Made gesture, *n* = 32	0.94 (0.80)	0.09 (0.39)	0.00 (0.00)
Did not make gesture, *n* = 53	0.58 (0.72)	0.02 (0.14)	0.08 (0.27)
		Posttest	
Made gesture, *n* = 32	1.53 (0.51)	0.47 (0.72)	0.34 (0.48)
Did not make gesture, *n* = 53	1.25 (0.78)	0.38 (0.60)	0.42 (0.69)
Sun motion gesture for sunrise		**Pretest**	
Made gesture, *n* = 37	0.81 (0.70)	0.03 (0.16)	0.03 (0.16)
Did not make gesture, *n* = 48	0.65 (0.81)	0.06 (0.32)	0.06 (0.24)
		Posttest	
Made gesture, *n* = 29	1.28 (0.70)	0.41 (0.68)	0.38 (0.62)
Did not make gesture, *n* = 56	1.39 (0.70)	0.41 (0.63)	0.39 (0.62)
Sun motion gesture for sunset		**Pretest**	
Made gesture, *n* = 41	0.85 (0.73)	0.05 (0.22)	0.05 (0.22)
Did not make gesture, *n* = 44	0.59 (0.79)	0.05 (0.30)	0.05 (0.21)
		Posttest	
Made gesture, *n* = 23	1.43 (0.74)	0.52 (0.73)	0.39 (0.58)
Did not make gesture, *n* = 62	1.32 (0.74)	0.37 (0.61)	0.39 (0.64)

N = 85. Standard deviations in parentheses.

[1]This sorting method does not distinguish participants who made no gesture and participants who made Sun motion gestures – both are assigned to the non-rotation-gesture group. We chose this method to preserve statistical power. Few participants made Sun motion gestures, and these few had similar verbal knowledge scores to the non-gesturers.

expressed about the same level of knowledge of sunrise as the non-gesturers, $t(83) = -0.021$, $p = 0.983$, and likewise for sunset, $t(83) = 0.095$, $p = 0.924$.

We obtained the same pattern of results when we sorted participants into Sun motion gesture and non-gesture groups on the basis of whether they made the Sun motion gesture when explaining *sunset* ($ps > 0.114$, see **Table 3** for mean knowledge components for gesturers and non-gesturers).

Do Pretest Rotation Gestures Predict Posttest Understanding?

To test whether gestures serve as a *window* into children's understanding, we tested whether pretest gestures predicted verbal knowledge at posttest. We focused on Earth rotation gestures, which signify the underlying cause of the day/night cycle. We conducted an ANCOVA with children's Earth rotation gestures predicting their posttest knowledge of the overall cause concept, controlling for children's knowledge of the overall cause concept at pretest. The analysis revealed that Earth rotation gestures at pretest *did not* predict verbal knowledge at posttest, $F(1, 84) = 0.365$, $p = 0.547$.

DISCUSSION

The current study examined whether and how children's spontaneous use of gesture relates to their explanations of the day/night cycle. Overall, participants used more Earth rotation gestures when explaining the overall cause of the day/night cycle – a topic for which both initial and eventual knowledge was relatively high. In fact, participants who made Earth rotation gestures tended to express *greater knowledge* than those who did not. Participants made more Sun motion gestures when explaining sunrise (though not for sunset, discussed further below) – topics for which initial and eventual knowledge was relatively low. Yet participants who made Sun motion gestures did not differ in verbally expressed knowledge from those who did not. As a whole, we did not observe a mismatch between participants' speech and gesture. Thus, in the present case, gestures generally *mirrored* the knowledge that children expressed verbally, rather than serving as a window into future knowledge gains.

By the posttest, the participants had been taught repeatedly about one essential fact: the Earth's rotation causes the change from daytime to nighttime. Nevertheless, participants were more likely to correctly explain the overall cause of the day/night cycle if they simultaneously expressed the Earth rotation gesture than if they did not. This finding supports the idea that children may require some physical bodily motion to fully encode abstract concepts – enactment theory (Matthews-Saugstad et al., 2016). The Earth's rotation, which can only be witnessed from outer space, may be better conceptualized with the enactment of a corresponding gesture. Gesture is distinct from other bodily movements in its unique ability to combine both physical and abstract elements (Goldin-Meadow, 2003, 2006). Whereas some scientific processes can be physically experienced with the body, other science concepts have abstract elements that cannot be experienced this way. For example, while a child may be able to experience gravity by jumping off the top of a box, it is impossible to physically experience the Earth's rotation. Gesture allows children to use their bodies to simulate such events (Goldin-Meadow, 2003, 2006; Novack and Goldin-Meadow, 2015) and could help them verbalize their understanding of spatial elements and relationships.

The participants also received repeated instruction about how the Earth's rotation causes sunrise and sunset. However, the posttest knowledge results show that participants tended to have less knowledge of these topics than for the overall cause of the day/night cycle. We suspect that participants struggled to connect their everyday experiences of witnessing an apparent Sun motion (i.e., the Sun going "up" in the morning) with the space-based perspective that the Sun is actually stationary, in part because of the misleading terms sun*rise* and sun*set*.

Interestingly, there was no relationship between participants' Sun motion gestures and their verbal-explanatory accuracy for either sunrise or sunset. Thus, whereas Earth rotation gestures signified understanding of the day/night cycle, Sun motion gestures did not clearly signify confusion about the cause of sunrise or sunset. This asymmetry may be due to the fact that the Sun motion gesture could reflect that a child is merely considering the Sun's apparent motion from an Earth-based perspective *or* that they actually believe that the Sun moves up and down – an intuitive but causally incorrect idea. In either case, the expression of the Sun motion gesture suggests that the child is focused on an Earth-based perspective of the day/night cycle, perhaps at the cost of forming a robust space-based representation that relates the Earth rotation to sunrise and sunset.

The current study has some important limitations. Firstly, children were never instructed to gesture in any way during the interviews. Because there was no manipulation of children's gestures, we cannot draw conclusions about causal links between gesture and understanding. An intriguing possibility is that encouraging Earth rotation gestures could enhance a student's understanding and their subsequent verbal explanations of the day/night cycle (see also Plummer, 2014). Indeed, people who are instructed to gesture when solving a spatially intense science problem, such as building a geologic block diagram, performed better on a subsequent spatial reasoning task than people who were prohibited from gesturing (Atit et al., 2015). Further research is needed to test this possibility in young children's astronomy learning.

Another limitation of the study is that there were small, but potentially important differences in the phrasing of certain interview questions. For the sunrise concept, children were asked, "Why does sunrise happen?" but for sunset, "What makes sunset happen?" The "why" phrasing for the sunrise question could have invited Earth-based and teleological responses such as "Sunrise happens so we wake up and go to school" or "Sunrise happens so a new day can start." The question about sunset may have invited more causal-mechanistic thinking, as the Earth's rotation is what *makes* sunset happen. So a small difference in wording could have affected how participants thought about the events in question. Future research should consider this issue when designing tests

to explore the effect of children's gestures on their knowledge of science concepts.

Finally, while the current results suggest gesture was a *mirror*, participants' actions during other parts of the interview could have revealed a window into knowledge. For example, children's modeling of the Earth and Sun motion using 3D balls could have served as a window into conceptual understanding. Future research could explore other components of the interview to capture a more complete picture.

Using gestures does not require expensive technology or extensive teacher training. Simply promoting the use of gesture in classrooms could be effective in improving children's understanding of highly spatial, abstract scientific concepts. Teachers can remind children to "use their hands" or demonstrate the kinds of gestures that students should use while learning concepts in class (e.g., Stieff et al., 2016). For example, when teaching about the day/night cycle, teachers might ask children to make an Earth rotation gesture with their hands while explaining to a peer how the Earth moves. Although we cannot say whether gesture *causes* deeper space science knowledge, the positive association between gesture and verbal knowledge suggests that gestures could be a contributing factor.

Beyond space science, gesture can support children's learning in other STEM domains. The geosciences (e.g., Atit et al., 2015) and organic chemistry (Stieff et al., 2016) offer a number of rich opportunities for incorporating gesture during instruction. Children also benefit from gestures during math lessons. Cook et al. (2008) found that children learned a mathematical concept better when they gestured during instruction rather than only speaking. Gesture may also be applied to other highly spatial, abstract domains, such as geometry, engineering, and architecture. Further research into direct applications of gesture for different scientific concepts will help inform recommendations for educators.

Gesture may serve an important role in the encoding of scientific information. In the current study, third graders who made Earth rotation gestures were more likely to verbally explain the day/night cycle than were those who did not produce this gesture. Gesturing to express highly spatial topics, such as the day/night cycle, may support children's knowledge acquisition. With additional research, gesture could be harnessed as a tool for instruction – a way to help people encode and express STEM concepts at a deeper level.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed for this study can be found in FigShare: https://doi.org/10.6084/m9.figshare.11592060.v1.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Committee on Human Subjects (IRB), College of the Holy Cross. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

CG and BJ conducted the statistical analyses. CG wrote the first draft of the manuscript. FA and BJ provided the critical revisions. CG and research assistants coded the video data. All authors conceptualized the idea for the project, contributed to the development and refinement of the gesture coding scheme, interpreted findings and outlined the manuscript, and approved the final version of the manuscript for submission.

FUNDING

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant No. R305A150228 to the College of the Holy Cross. The opinions expressed are those of the authors and do not represent the views of the Institute of Education Sciences or the U.S. Department of Education. This research was also supported by the Charles S. Weiss undergraduate summer research fellowship from the College of the Holy Cross.

ACKNOWLEDGMENTS

We thank Natalie Evans, Amanda McCarthy, and Victoria Jackson for developing the coding scheme and coding interviews, as well as all research assistants who assisted with data collection. We are grateful to the schools, teachers, parents, and students who took part in the research.

REFERENCES

Alibali, M. W. (2005). Gesture in spatial cognition: expressing, communicating, and thinking about spatial information. *Spat. Cogn. Comput.* 5, 307–331. doi: 10.1207/s15427633scc0504_2

Alibali, M. W., and Kita, S. (2010). Gesture highlights perceptually present information for speakers. *Gesture* 10, 3–28. doi: 10.1075/gest.10.1.02ali

Alibali, M. W., and Nathan, M. J. (2018). "Embodied cognition in learning and teaching: Action, observation, and imagination. In F. Fischer, Hmelo-Silver," in *International Handbook of the Learning Sciences*, eds C. E. Hmelo-Silver, S. R. Goldman, and P. Reimann (New York, NY: Routledge).

Alibali, M. W., Spencer, R. C., Know, L., and Kita, S. (2011). Spontaneous gesture influences strategy choices in problem solving. *Psychol. Sci.* 22, 1138–1144. doi: 10.1177/0956797611417722

Alibali, M. W., Young, A. G., Crooks, N. M., Yeo, A., Wolfgram, M. S., Ledesma, I. M., et al. (2013). Students learn more when their teacher has learned to gesture effectively. *Gesture* 13, 201–233. doi: 10.1075/gest.13.2.05ali

Atit, K., Gagnier, K., and Shipley, T. F. (2015). Student gestures aid penetrative thinking. *J. Geosci. Educ.* 63, 66–72. doi: 10.5408/14-008.1

Carr, M., Alexeev, N., Lu, W., Barned, N., Horan, E., and Reed, A. (2017). The development of spatial skills in elementary school students. *Child Dev.* 89, 446–460. doi: 10.1111/cdev.12753

Church, R. B., and Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition* 23, 43–71. doi: 10.1016/0010-0277(86)90053-3

Cook, S. W., Mitchell, Z., and Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition* 106, 1047–1058. doi: 10.1016/j.cognition.2007.04.010

Field, A. (2013). *Discovering Statistics Using IMB SPSS Statistics*, 4th Edn. Thousand Oakes, CA: Sage.

Goldin-Meadow, S. (2003). *Hearing Gesture: How Our Hands Help us Think*. Cambridge, MA: Harvard University Press.

Goldin-Meadow, S. (2005). *Hearing Gesture: How Our Hands Help us Think*. Cambridge, MA: Harvard University Press.

Goldin-Meadow, S. (2006). Talking and thinking with our hands. *Curr. Dir. Psychol. Sci.* 15, 34–39. doi: 10.111/j.0963-7214.2006.00402.x

Goldin-Meadow, S. (2015). From action to abstraction: gesture as a mechanism of change. *Dev. Rev.* 38, 167–184. doi: 10.1016/j.dr.2015.07.007

Heywood, D., Parker, J., and Rowlands, M. (2013). Exploring the visuospatial challenge of learning about day and night and the sun's path. *Sci. Educ.* 97, 772–796. doi: 10.1002/sce.21071

Howell, D. C. (2002). *Statistical Methods for Psychology*, 5th. Edn. Belmont, CA: Duxbury.

Iverson, J. M., and Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychol. Sci.* 16, 367–371. doi: 10.1111/j.0956-7976.2005.01542

Jee, B. D., and Anggoro, F. K. (2019). Relational scaffolding enhances children's understanding of scientific models. *Psychol. Sci.* 30, 1287–1302. doi: 10.1177/0956797619864601

Lee, J., and Bednarz, R. (2012). Components of spatial thinking: evidence from a spatial thinking ability test. *J. Geogr.* 111, 15–26. doi: 10.1016/j.sbspro.2011.07.048

Matthews-Saugstad, K. M., Erik, P., and Kelty-Stephen, D. G. (2016). Gesturing more diminishes recall of abstract words when gesture is allowed and concrete words when it is taboo. *Q. J. Exp. Psychol.* 70, 1099–1105. doi: 10.1080/17470218.2016.1263997

Newcombe, N. S., and Shipley, T. F. (2015). "Thinking about spatial thinking: New typology, new assessments," in *Studying Visual and Spatial Reasoning for Design Creativity*, ed. J. S. Gero (Berlin: Springer), 179–192. doi: 10.1007/978-94-017-9297-4_10

Novack, M., and Goldin-Meadow, S. (2015). Learning from gesture: How our hands change our minds. *Educ. Psychol. Rev.* 27, 405–412. doi: 10.1007/s10648-015-9325-3

Özçalışkan, S., Gentner, D., and Goldin-Meadow, S. (2014). Do iconic gestures pave the way for children's early verbs? *Appl. Psycholinguist.* 35, 1143–1162. doi: 10.1017/S014271612000720

Plummer, J. D. (2014). Spatial thinking as the dimension of progress in an astronomy learning progression. *Stud. Sci. Educ.* 50, 1–45. doi: 10.1080/03057267.2013.869039

Plummer, J. D., Bower, C. A., and Liben, L. S. (2016). The role of perspective taking in how children connect reference frames when explaining astronomical phenomena. *Int. J. Sci. Educ.* 38, 345–365. doi: 10.1080/09500693.2016.1140921

Rusticus, S. A., and Lovato, C. (2014). Impact of sample size and variability on the power and type I error rates of equivalence tests: a simulation study. *Pract. Assess. Res. Eval.* 19, 1–10. doi: 10.7275/4s9m-4e81

Shtulman, A. (2017). *Scienceblind: Why Our Intuitive Theories About the World Are so Often Wrong*. New York, NY: Basic Books.

Stieff, M., Lira, M. E., and Scopelitis, S. A. (2016). Gesture supports spatial thinking in STEM. *Cogn. Instr.* 34, 80–99. doi: 10.1080/07370008.2016.1145122

Uttal, D. H., and Cohen, C. A. (2012). Spatial thinking and STEM education: when, why and how. *Psychol. Learn. Motivat.* 57, 147–181. doi: 10.1016/B978-0-12-394293-7.00004-2

Wai, J., Lubinski, D., and Benbow, C. P. (2009). Spatial ability for STEM domains: aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Am. Psychol. Assoc.* 101, 817–835. doi: 10.1037/a0016127

Waters, G. M., and Beck, S. R. (2015). Verbal information hinders young children's ability to gain modality specific knowledge. *Infant Child Dev.* 24, 538–548. doi: 10.1002/icd.1901

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Gaudreau, Anggoro and Jee. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 10 June 2020
doi: 10.3389/fpsyg.2020.01047

Parental Pre-knowledge Enhances Guidance During Inquiry-Based Family Learning in a Museum Context: An Individual Differences Perspective

Rooske K. Franse[1,2], Tessa J. P. Van Schijndel[3] and Maartje E. J. Raijmakers[1,2,4]*

[1] Department of Psychology, University of Amsterdam, Amsterdam, Netherlands, [2] NEMO Science Museum, Amsterdam, Netherlands, [3] Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, Netherlands, [4] Department of Education Sciences, LEARN!, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

OPEN ACCESS

Edited by:
Christopher Osterhaus, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Timo Reuter, University of Koblenz and Landau, Germany
Josh Gutwill, Exploratorium, United States

***Correspondence:**
Rooske K. Franse
r.k.franse@uva.nl

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *27 January 2020*
Accepted: *27 April 2020*
Published: *10 June 2020*

Citation:
Franse RK, Van Schijndel TJP and Raijmakers MEJ (2020) Parental Pre-knowledge Enhances Guidance During Inquiry-Based Family Learning in a Museum Context: An Individual Differences Perspective. Front. Psychol. 11:1047. doi: 10.3389/fpsyg.2020.01047

Effective interaction and inquiry are an essential source for children's learning about science in an informal context. This study investigated the effect of parental pre-knowledge on parent–child interactions (manipulations, parent talk, and child talk) during an inquiry activity in NEMO Science Museum in Amsterdam. The sample included 105 parent–child dyads (mean children's age = 10.0 years). Half of the couples were randomly assigned to the experimental group in which, without the child's knowledge, the parent was shown the task's solution prior to the inquiry activity. Results show that parental pre-knowledge affected the way parents interacted and inquired with their child. Compared to parents without pre-knowledge, parents with pre-knowledge inquired longer, posed more open-ended *wh*-questions and closed questions, and less often interpreted results. Children of parents with pre-knowledge more often described evidence and interpreted results, more often manipulated alone, and solved the task more accurately. These results indicate that parental pre-knowledge brings about parents' scaffolding behavior. In addition, it was studied how individual differences of parents and children relate to parent–child interaction. Results show that children's self-reported inquiry attitude was related to their conversation during inquiry, such that they asked fewer closed questions and more open-ended questions. Children's gender affected the cooperation between parent and child, parents more often manipulated together with boys than with girls, and girls more often manipulated alone. Fathers with pre-knowledge, but not mothers, let their child manipulate more by oneself than fathers without pre-knowledge. This study shows that more knowledge about an exhibit improves a parent's scaffolding behavior in a science museum. Results are discussed in the context of museum practice.

Keywords: *wh*-questions, parent–child interaction, pre-knowledge, individual differences, inquiry-based learning, museum context

INTRODUCTION

Science museums offer families opportunities to learn from and about everyday science mostly by inquiry-provoking activities, which are an important means to learn about these phenomena (National Research Council, 2009). Typical for inquiry is the *gathering of evidence* by manipulating materials and observing effects, and the *interpreting of evidence* by discussing the observed effects, linking observations to pre-knowledge, and weighing the quality of gathered evidence (Schauble, 1990, 1996; Legare et al., 2017). Inquiry activities can provide parents and children with opportunities to gain insight into specific phenomena (Fender and Crowley, 2007), to develop and practice inquiry skills (Gutwill and Allen, 2010), and to experience interest in science (Crowley and Jacobs, 2002).

How parents and children inquire and what they can learn through inquiry in the museum context is of interest for developmental and educational researchers (Sobel and Jipson, 2015), as well as for museum practitioners (Allen, 2002, 2004). Parent–child interaction during inquiry activities in the museum has been studied by focusing at different aspects of behavior (Haden, 2010; Legare et al., 2017): verbal behavior (e.g., Callanan and Jipson, 2001; Gauvain, 2001; Callanan and Valle, 2008; Benjamin et al., 2010; Kisiel et al., 2012; Luce et al., 2013; Tenenbaum and Hohenstein, 2016) and non-verbal behavior (e.g., Crowley et al., 2001a; Van Schijndel et al., 2010; Willard et al., 2019) of parent, child, or parent–child dyads. Parent–child interaction during inquiry activities has been studied in different content areas, including physics (e.g., Crowley et al., 2001a), engineering (e.g., Benjamin et al., 2010), and biology (e.g., Eberbach and Crowley, 2017).

Inquiry activities offer many opportunities for learning, but does not automatically result in new knowledge or skills. Research into open-ended inquiry in the school context has demonstrated that inquiry is not always effective for concept learning and that teacher guidance (e.g., scaffolding) substantially contributes to the learning outcomes (Alfieri et al., 2011). By scaffolding, teachers temporarily bridge the gap between a learning task and children's current abilities (Wood et al., 1976). Types of scaffolding that are used in inquiry learning are modeling, questioning, giving hints, instructing, or explaining (Van de Pol et al., 2010). Teachers can, for example, think out loud, or model how to ask questions (Rosenshine, 2009) and, by doing so, reduce the difficulty of an ill-structured task. This type of teacher support has a positive effect on children's knowledge and skill acquisition if it is in line with the child's pre-knowledge (Van de Pol et al., 2010; Alfieri et al., 2011). In the museum context, parents could guide their children during inquiry activities by giving individual attention and support (Crowley and Callanan, 1998; Ash, 2004; Pattison and Dierking, 2013). Research has shown that children inquire longer and on a deeper level (e.g., hypothesis-driven) if accompanied by their parents compared to inquiring alone or with peers (Gleason and Schauble, 1999; Crowley et al., 2001a). However, in general, parents are not professional teachers (Schauble et al., 2002), and it has also been demonstrated that parents can miss out on opportunities to support their children's learning potential (Gleason and Schauble, 1999; Palmquist and Crowley, 2007; Eberbach and Crowley, 2017). For example, parents sometimes lack specific content knowledge that could enrich the verbal interaction between parent and child (Knutson and Crowley, 2010), or assume that the child's understanding is similar to their own when interpreting evidence (Gleason and Schauble, 1999).

Parent–child interaction during inquiry activities is considered to be a collaborative and dynamic process of exploring and explaining (Legare et al., 2017). The process is described as collaborative, because both parent and child add to the learning situation by their behavior and talk, while interacting with each other (Callanan and Oakes, 1992; Gleason and Schauble, 1999). The process of parent–child interaction is described as dynamic, because the processes of *gathering evidence* through inquiry and *interpreting evidence* by drawing on prior experiences and knowledge (Siegel et al., 2007) mutually influence each other.

Within this collaborative and dynamic learning process, parent and child differences in knowledge, reasoning skills, and interest will result in opportunities for parents to scaffold their children's learning (Wood et al., 1976; Alfieri et al., 2011). However, also for parents, content presented in the museum is sometimes new or complex. Science museums often present exhibits covering a multitude of phenomena from different content areas. Hence, it is not possible for parents to have a good understanding of all of these. This means that parents during joint inquiry often take on not only the role of facilitator of the child's learning process but also the role of learner (Siegel et al., 2007; Falk, 2009). The present study is aimed at improving our understanding of how parents' conceptual pre-knowledge affects parent–child interaction during inquiry activities in the museum.

Pre-knowledge and Parent–Child Interaction During Inquiry Activities

Previous research in both formal and informal learning contexts has shown that pre-knowledge affects the way people inquire (Klahr and Dunbar, 1988; Trumbull et al., 2005) and interact (Palmquist and Crowley, 2007; Eberbach and Crowley, 2017). In a formal learning context, a lack of pre-knowledge has been shown to impede the way people experiment and make observations (Klahr and Dunbar, 1988; Trumbull et al., 2005).

In an informal learning context, two correlational studies (Palmquist and Crowley, 2007; Eberbach and Crowley, 2017), and only one experimental study (Benjamin et al., 2010) investigated the relation of *parents'*, *children's*, and *dyads'* pre-knowledge with parent–child interaction. Eberbach and Crowley (2017) demonstrated in the context of a botanical garden, with 6- to 10-year-old children, that parental pre-knowledge is related to parent–child verbal interaction and parental guidance style. Compared to parents who knew less, parents who knew more about pollination more often talked about pollination, and more often asked their children content-related open-ended *wh*-questions during the garden visit. Open-ended *wh*-questions

start with, for example, What, Why, or How, and aim at stimulating dialogue (Haden, 2010) and focusing the attention on relevant aspects for remembering and learning (Falk and Dierking, 2000; Leinhardt and Knutson, 2004; Benjamin et al., 2010). Palmquist and Crowley (2007) studied the relation of children's pre-knowledge and parent–child interaction in the context of a Dinosaur exhibition, with 5- and 7-year-old children. It was demonstrated that children's pre-knowledge about dinosaurs and paleontology was related to parent talk: higher amounts of talk was observed with parents of novice children compared to parents of expert children. Moreover, based on exploratory observations, children's pre-knowledge appeared to be associated with parents' guidance. Parents of expert children seemed "testers" of the child's knowledge; they, for example, asked questions that encouraged children to present their knowledge about dinosaurs. Parents of novice children, on the other hand, seemed "teachers," who supported the child's learning process and inquired along with their child, by, for example, interpreting the information that was presented in the exhibition. These two studies showed positive relations between parents' or children's domain-specific pre-knowledge and parents' domain-general guidance. However, as the studies were correlational in nature, observed differences in parental scaffolding behaviors could possibly be explained by other person characteristics than pre-knowledge. For example, parents' attitude toward learning has shown to be related to both parental knowledge acquisition and parental guidance (Sigel, 1998; Sigel and McGillicuddy-De Lisi, 2002; Ricco and Rodriguez, 2006). Benjamin et al. (2010) investigated, in a study with an experimental design, the effect of parents' and children's knowledge on parent–child interaction in the context of a Building exhibition, with 4- to 8-year-old children. It was found that by receiving domain-specific knowledge prior to visiting the exhibition, parents' and children's content-related talk and building behavior improved: dyads' conversations (*wh*-questions and associations) were more often domain-specific and their buildings were sturdier. Pre-knowledge, however, did not affect parental scaffolding: the total amount of *wh*-questions and associations did not increase. Possibly, the lack of effect on parental scaffolding is explained by the fact that, in this study, parent and child received the same content-related information. In the current study, we will therefore investigate the effect of parental pre-knowledge on parental scaffolding during inquiry activities in an experimental design in a museum context. We focus on 8- to 12-year-old children. Evidently, apart from parental pre-knowledge, more person characteristics of both the parent and the child are possibly relevant for parent–child interaction during inquiry, such as age (e.g., Kuhn et al., 1988; Klahr et al., 1993; Schauble, 1996), educational level (Callanan and Jipson, 2001), epistemic beliefs about learning (Ricco and Rodriguez, 2006), executive and cognitive functioning (Kirschner et al., 2006; Brigham et al., 2011; Watt et al., 2014; Zweers et al., 2019), and motivation and interest (Tomlinson et al., 2003; Vansteenkiste et al., 2004). Therefore, we will include person characteristics of parent and child in the current study. Below, we will briefly introduce research related to the impact of person characteristics on parent–child interaction in the museum context.

Person Characteristics and Parent–Child Interaction During Inquiry Activities

Person characteristics, such as children's and parents' age, gender, and educational level, are reported in most museum research to give insight into the population that is studied. Some research, however, also studied how parent (e.g., Siegel et al., 2007; Tare et al., 2011; Nadelson, 2013) or child characteristics (e.g., Geerdts et al., 2015) are related to parent–child interaction. This research shows that individual differences are large, and their impact on learning and behavior in the museum might be substantial.

Gender

With regard to gender, it has been found that parents interact differently with boys and girls in the museum context (Crowley et al., 2001b; Siegel et al., 2007; Luce et al., 2013). Parents gave more causal explanations of science content to boys (Crowley et al., 2001b), made more absolutist statements such as claims and facts to boys (Luce et al., 2013), and behaved more collaboratively with boys (Siegel et al., 2007). Additionally, fathers and mothers have been shown to interact differently with their children in museums (Benjamin et al., 2010; Nadelson, 2013; Van Schijndel and Raijmakers, 2016). For example, father–child dyads played longer in a construction exhibition (Benjamin et al., 2010), and mothers gave more *causal* explanations (Van Schijndel and Raijmakers, 2016).

Interest, Motivation, and Attitude

Science interest is seen as a multi-component construct, where behavior, enjoyment, knowledge components, values, and motivational aspects mutually influence each other (Ainley, 2017; Sachisthal et al., 2018). Parents' interest in science has been shown to be related to parent–child interaction at exhibits: parent and child engaged with more exhibits if parents had a positive science attitude (Szechter and Carey, 2009), and parent and child spent more time at exhibits if the science topic of an exhibit was of interest to parents (Tare et al., 2011). Besides science interest, other motivational aspects such as parents' agenda to visit a museum have also been shown to be related to parent–child interaction. For example, if parents' motivation to visit the museum was educational, parents were more involved in the child's learning at an exhibit (Tare et al., 2011).

Age

One would expect children's inquiry behaviors to be age-related (Kuhn et al., 1988; Klahr et al., 1993; Schauble, 1996). However, museum research did not show age-related differences in children's content-related talk (Geerdts et al., 2015, 3- to 8-year-olds; Marcus et al., 2018, 4- to 8-year-olds) or manipulations (Fender and Crowley, 2007, 5- to 7-year-olds) during inquiry. In comparable age ranges, age-related differences were found in children's conceptual understanding of the exhibit (Fender and Crowley, 2007). In addition, parents' behavior was found to be related to children's age (Geerdts et al., 2015; Marcus et al., 2018): compared to school-aged children, parents more often talked with preschoolers about non-observable characteristics (Geerdts et al., 2015), and science processes, technology, and

engineering (Marcus et al., 2018). Other studies on parent–child conversations (Jipson and Callanan, 2003, 3- to 5-year-olds; Tenenbaum and Leaper, 2003, 11- to 13-year-olds) or manipulations (Fender and Crowley, 2007, 5- to 7-year-olds) did not find differences in parents' behavior in relation to children's age.

Educational Level

Parents' schooling has shown to be related to how parents interact with their children in the museum context (Siegel et al., 2007; Szechter and Carey, 2009). Parents' educational level and science museum experience was positively associated with time spent inquiring at an exhibit, and with the frequency that dyads linked the inquiry to prior experiences (Szechter and Carey, 2009). In addition, it was found that in a science museum context, higher educated parents were more directive than lower educated parents (Siegel et al., 2007).

Working Memory

Cognitive and executive functioning is important for learning (Kirschner et al., 2006), especially in a museum environment with a lot of distraction and open discovery tasks (Allen, 2004). An overloaded working memory can affect children's learning experiences (Rosenshine, 2009).

Beliefs About Learning

Additionally, parents' epistemological beliefs about learning might be relevant for parent–child interaction in the museum. Parents' beliefs about learning have shown to be related to parental guidance (Sigel, 1998; Sigel and McGillicuddy-De Lisi, 2002; Ricco and Rodriguez, 2006).

Current Study

The current study aims at a better understanding of parent–child interaction during inquiry activities, a type of activity that is at the core of the science museum experience. Parental pre-knowledge appears to play a role in parent–child interaction during inquiry, but most previous insights have stemmed from correlational research. We present a study with an experimental design, in which we manipulated parental pre-knowledge, addressing two research questions:

- How does parental pre-knowledge affect parent–child interaction during an inquiry activity in the museum? (RQ-1).
- How do person characteristics (i.e., parents' gender, educational level, science interest and beliefs about learning, and children's age, gender, working memory, enjoyment in science lessons, and inquiry attitude) affect parent–child interaction during an inquiry activity, and the possible relation of parental pre-knowledge and parent–child interaction? (RQ-2).

Our hypotheses are based on the idea that parents with domain-specific pre-knowledge about the phenomenon of inquiry do not have an urge for information, and therefore have the opportunity to scaffold their child's learning process.

Operationalization and Hypotheses

To study the possible causal effect of parental pre-knowledge on parent–child interaction (RQ-1), a randomized controlled trial was designed, with two conditions: A control condition without pre-knowledge and an experimental condition, in which *parents* received conceptual knowledge about the phenomenon to inquire in the inquiry activity.

To be able to control *parental* pre-knowledge, a so-called black-box was used as the object of inquiry, an object that does not allow one to see from the outside what is going on inside (Miller, 2014). An important characteristic of a black-box is that no physical laws are applicable (e.g., shadow size, buoyancy, magnetism), and therefore participants cannot have pre-knowledge about the problem to be solved. This way, one can experimentally control parental pre-knowledge. The black-box used in the current study consisted of a wooden box with four holes from which four rope ends protruded. How the ropes were entangled inside the box could not be observed from the outside; however, it could be discovered by manipulating the ropes. Parents in the pre-knowledge condition were shown the entangled ropes in the inside of the box prior to the inquiry activity, without their child being aware of this disclosure. The black-box was offered in a separate room, the Research and Development lab, at the museum floor of NEMO Science Museum.

To study how person characteristics affect parent–child interaction (RQ-2), children performed a task (cognitive abilities) prior to the black-box inquiry, and parents and children filled out questionnaires afterward. Children aged between 8 and 12 participated in the study. In this age range, children can already contribute to conversations about inquiry-related topics (Chinn and Malhotra, 2002).

Parent–child interaction was measured by observing behavior and talk during inquiring at the black-box. Behavior during inquiry consisted of pulling one or more ropes, and observing what the causal effect of this manipulation was. This could be the parent's, the child's (solitary), or cooperative behavior. We expect that parents without pre-knowledge have an urge for information and are primarily focused on finding out for themselves how the problem can be solved. We therefore expect them to perform more manipulations of the ropes by themselves, compared to parents with pre-knowledge. We expect that parents with pre-knowledge shift from a role of learning along with their child toward taking the role of facilitating the child's learning (Siegel et al., 2007; Falk, 2009). We therefore expect that they less often manipulate the ropes by themselves.

We quantified parents' and children's talk in terms of elements of scientific reasoning, such as formulating hypotheses and interpreting results, and type of explanatory talk, such as asking open-ended *wh*-questions, asking closed questions, describing evidence, and giving directions (Crowley et al., 2001a; Fender and Crowley, 2007; Gutwill and Allen, 2010; Van Schijndel and Raijmakers, 2016). We expect that pre-knowledge will affect the parental talk, such that parents with pre-knowledge will better facilitate children's learning process. That is, we expect that parents with pre-knowledge talk more and ask more open-ended

wh-questions to encourage children to formulate hypotheses and interpret results (Crowley and Jacobs, 2002).

MATERIALS AND METHODS

Participants

One hundred sixteen parent–child dyads visiting NEMO Science Museum in Amsterdam participated in the study. Eleven dyads were excluded from analyses, 1 for retracting permission and 10 for technical problems. The final sample included 105 dyads consisting of an adult (P; M_{age} = 43.18, SD = 4.92; 50 male, 55 female), and a child (C; M_{age} = 9.96, SD = 1.38; 51 boys, 54 girls). To sketch a profile of the participating dyads: in all cases, the adult was the caretaker of the child. Parents were relatively highly educated (19% Graduate, 45% Bachelor, and 36% Up to Bachelor's), and moderately interested in science. For example, half of the parents reported reading the science supplement of an (internet) newspaper weekly to monthly. Almost all parents had visited half-yearly to annual a science or natural history museum, and over half of the parents had watched or listened weekly to monthly to a science program on the radio or television. Mostly, parents believed that children learn by experimenting, reasoning, and drawing conclusions, trial and error, gaining success experiences, and receiving positive feedback from their parents (see *Section* "Measures for Person Characteristics"). Children were highly engaged in science, which is to be expected in a science museum.

Procedure and Study Design

The study was conducted in the museum during spring break 2016. Families with children (8- to 12-year-olds) were approached and asked if they wanted to participate in a scientific study. If families agreed, they were welcomed in a research room where research assistants explained the procedure. The parent completed written consent forms for him- or herself and the child. Parent–child dyads were randomly assigned to a with pre-knowledge experimental condition (E, N = 54) or a without pre-knowledge control condition (C, N = 51). For both conditions, the experimental session took about 20 min and included an inquiry task that dyads could play with as long as they wanted, with a maximum of 5 min. Prior to the inquiry task and without the child being aware of it, parents in the with pre-knowledge experimental condition were shown the inside of the box. While inquiring, parent–child interaction was video recorded. When finished, parent and child were asked to make a drawing, each separately, as a measure for learning through inquiry. Furthermore, the experimental session consisted of measures to characterize the population. After the inquiry activity, parents filled out a questionnaire consisting of *Background* questions (age, gender, and educational level), *Science interest* statements, and *Beliefs about learning processes* statements. Children, prior to the inquiry task, performed a visual spatial *Working memory* task and, after inquiring, filled out a questionnaire consisting of questions about *Enjoyment in science lessons*, *Attitude to inquiry*, and *Enjoyment in science*.

Materials

Inquiry Task

To study the effect of parental pre-knowledge on parent–child interaction, an inquiry task was selected that, on the one hand, encouraged hypothesis-driven inquiry and, on the other hand, provided a challenge for which participants have no specific pre-knowledge, but for which pre-knowledge could be given in a quick and unambiguous way. Black-box tasks met those criteria (Lederman and Abd-El-Khalick, 1998). The black-box used in the study consisted of a wooden box (25 × 15 × 10 cm) with four holes, two ropes, one fabric ring, and a padlock (see also **Figure 1**). Inside the box, both ropes run through the fabric ring. Each rope end had a unique color (blue and red for rope 1, green and yellow for rope 2), and protruded through one of four box holes. When the box was closed, only the rope ends were visible, and how the ropes were entangled inside could not be observed. The way the ropes were entangled caused a complex movement pattern. This movement pattern was partly caused by the fabric ring that is not at a fixed position in the box but can also move. When someone pulled a rope, another rope end was pulled in (or multiple rope ends were pulled in). For example:

- If a participant pulls one rope (e.g., red), then the other three ropes (yellow, green, and blue) will be pulled in. How much these three ropes are pulled in depends on how tightly the participant pulls and the current position of the fabric ring in the black-box.
- If a participant pulls two ropes on the short side of the box (e.g., blue and red), then the other two ropes (green and yellow) are pulled in until the participant can't pull the rope ends any further.
- If a participant pulls two ropes on the long side of the box (e.g., blue and green), then the other two ropes (red and yellow) are pulled in without restriction, which is that the red and yellow rope ends could disappear into the box (this, however, never happened).

Parent–child dyads were presented with the closed box and were asked to "inquire how the ropes are running on the inside." Families were free to follow their own approach (e.g., pulling one rope at a time, or pulling two ropes simultaneously) and to inquire as long as they wanted, with a maximum of 5 min.

Task-Related Information (Pre-knowledge)

Parents in the experimental condition, but not in the control condition, were invited to peek into the black-box to observe how the ropes were entangled, just before the inquiry task started and without the child being aware of it.

Coding Approach

Parent and child's inquiry process was recorded on video. The final scoring was based on transcripts of these recordings (in CLAN: MacWhinney and Snow, 1990). A transcript was first broken down into speech segments. A segment ended if the parent or child were taking turns or ended after a natural silence. That is, silences were included in the preceding speech segment. Parents' and children's manipulations during inquiry

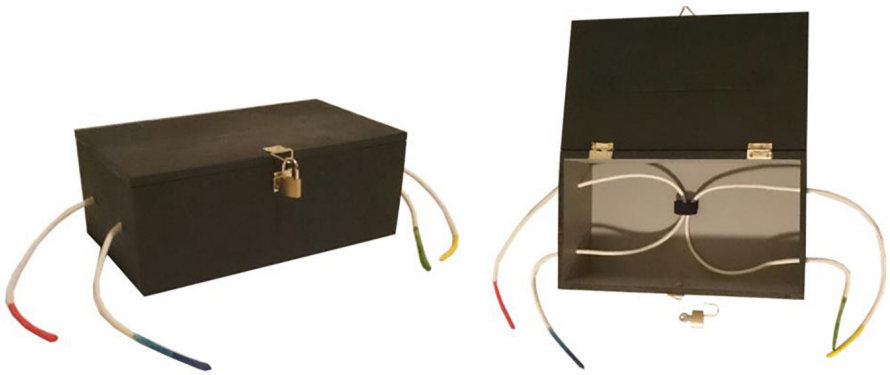


FIGURE 1 | Black-box inquiry task. **(A)** Closed black-box, as presented to parent–child dyads during inquiry. Four rope ends, with unique colors (red, blue, green, and yellow) stick out. The box is sealed by a padlock. **(B)** Opened black-box, as presented to parents in the pre-knowledge experimental condition, prior to the inquiry task. Parents could observe two ropes, one with a red and blue rope end and another with a green and yellow rope end interconnected through a fabric ring.

were scored by classifying each speech segment using a five-subscale coding instrument. The highest score was used for manipulations during a speech segment. In case a manipulation was continuing over multiple speech segments, it was only scored once. These subscales were: No manipulation (M_0), Child manipulates alone (M_C), Parent manipulates alone (M_P), Parent and Child manipulate in parallel ($M_{C/P}$), and Parent and Child manipulate together ($M_{C\&P}$). The inter-observer reliability for manipulations (two observers, 20% of the data) was found to be "substantial" (Landis and Koch, 1977): the percentage agreement was 85% and kappa was 0.77 ($p < 0.001$), 95% CI (0.73, 0.80). Frequencies of the five manipulation types were used as outcome variables in further analyses. Parents' (P) and children's (C) individual contribution to the conversations was scored by classifying each speech segment using a seven-subscale coding instrument distinguishing six different types of inquiry and guidance (Crowley et al., 2001a; Zimmerman, 2005; Fender and Crowley, 2007; Gutwill and Allen, 2010; Van Schijndel and Raijmakers, 2016). These types were as follows: Asking open-ended *wh*-questions (C1-C and C1-P for children and parents respectively; example: "Why does that rope move?"), Asking closed questions (C2-C and C2-P; example: "Is this rope attached to that rope?"), Describing evidence (C3-C and C3-P; example: "If I pull the red one then the blue one moves, but not the green one."), Interpreting results (C4-C and C4-P; example: "The blue rope only pulls the red rope, therefore, they belong together."), Giving direction (C5-C and C5-P; example: "Go ahead, just pull a rope."), and Formulating hypotheses (C6-C and C6-P; example: "I expect that these four ends are actually two separate ropes."). The seventh subscale (C7-C and C7-P) contained all unclassifiable comments such as expressing emotions (e.g., "This is really fun to do"). The inter-observer reliability for conversations (two observers, 20% of the data) was found to be "almost perfect" (Landis and Koch, 1977): the percentage agreement was 94% and kappa was 0.93 ($p < 0.001$), 95% CI (0.91, 0.95). The remaining comments category (C7-C and C7-P) was not included in the analyses. As a learning outcome variable, parents' and children's drawings were classified in four categories: Incorrect ropes and incorrect connection (K1), Correct ropes but incorrect fixed connection (K2), Correct ropes but incorrect loose connection (K3), and Correct ropes and correct loose connection (K4). Holding time, the number of minutes played, was used as a first explorative outcome variable to describe parent–child interaction.

To sum up, parent–child interaction is in the current study described by 17 dependent variables. That is, five manipulations variables, six parent talk variables and six child talk variables. Learning (knowledge gain) is described by four dependent variables.

Measures for Person Characteristics

Working memory (child)

Children's visual spatial working memory was tested using the Chessboard Task (Dovis et al., 2012). This task assesses children's ability to both maintain and manipulate visual–spatial information, and is based on the Corsi Block Tapping Task (Corsi, 1972), and the subtest Letter–Number Sequencing from the Wechsler Adult Intelligence Scale (WAIS, Wechsler, 1958). Children played for a maximum of 7 min, and on average performed 20 trials. As a measure of the child's working memory, the longest sequence achieved during 7 min of play was reported and used in further analyses. The working memory scores are relative scores and are solely used to compare differences between parent–child dyads within the current study.

Enjoyment in science lessons (child) and Attitude to inquiry (child)

Subscales *Enjoyment in science lessons* ($\alpha = 0.91$, 10 items) and *Attitude to inquiry* ($\alpha = 0.81$, 10 items) of the Test of Science-related attitudes (TOSRA), a measure to distinguish science-related attitudes among secondary school students (Fraser, 1981), were translated to Dutch and adjusted to primary school wording. Children rated their agreement with statements on a 5-point-Likert scale (1 = strongly disagree to 5 = strongly agree). Example questions are as follows: "Science is one of the most interesting school subjects" (*Enjoyment in science lessons*

sub-scale), and "I would rather solve a problem by doing an experiment than be told the answer" (*Attitude to inquiry* sub-scale). A forced two-factor analysis under Varimax rotation of the Dutch questionnaire (i.e., the translated and adjusted subscales), resulted in two process factors that contained the same items and comparable reliabilities as reported by Fraser (1981): *Enjoyment in science lessons* with α = 0.91 (cf. Fraser, α = 0.91) and *Attitude to inquiry* with α = 0.74 (cf. Fraser, α = 0.81), explaining 30 and 15% of variance, respectively. Sum scores of *Enjoyment in science lessons* and sum scores of *Attitude to inquiry* are reported and used in further analyses.

Additional measures for person characteristics, not used in analyses

Parents' *Beliefs about learning processes* (i.e., learning as active or passive process) were evaluated using a 16-statement survey (How Children Learn Inventory; Ricco and Rodriguez, 2006). However, the reliabilities of the two sets of statements, α = 0.38 for learning as an active process and α = 0.45 for learning as a passive process, were insufficient to use variables based on these sets in further analyses. Explorative factor analysis resulted in one scale with 10 statements and α = 0.62, explaining 16% of the variance. In addition to children's *Enjoyment in science lessons*, also children's not school-related *Science enjoyment* was evaluated using a subscale of the Dutch science and technology attitude instrument for primary school pupils (Walma van der Molen et al., 2007). Children's responses on the VTB and TOSRA subscales were found to be significantly related, $r = 0.51$, $p < 0.001$; therefore, only the results of the subscale with the highest reliability were included in the analyses (i.e., TOSRA *Enjoyment in science lessons*).

To sum up, in the analyses, seven independent variables will be used to describe parents' and children's person characteristics, that is, five child characteristics—age, gender, working memory, enjoyment in science lessons, and inquiry attitude—and two parent characteristics—gender and educational level.

Analysis Approach

To study the effect of parental pre-knowledge, person characteristics, and possible interactions between parental pre-knowledge and person characteristics on parent–child interaction, three MANCOVAs were performed, one for each aspect of parent–child interaction: manipulations, parent talk, and child talk. To further study the relationship between the dependent variables of each parent–child interaction aspect, follow-up analyses were performed using univariate ANOVAs (Field, 2009). From these analyses, we learned how the behavioral measures (e.g., the six parent talk categories) play together in different conditions (RQ-1; with and without parental pre-knowledge), or with different person characteristics (RQ-2; e.g., gender child). Levene's Test of Equality of Error Variances was violated for six outcome variables (M_0, $M_{P\&C}$, C1-P, C6-P, C1-C, and C2-C); therefore, we choose Hotelling's *T* as test statistics (Field, 2009); nevertheless, robustness of the *F* statistic seems to be warranted because of equal group sizes of pre-knowledge (without = 51, with = 54), inquiry attitude (low = 55, high = 50), and gender child (boys = 51, girls = 54) groups, also see **Table 2** (Blanca et al., 2018). To test the robustness of results, we performed non-parametric tests (Mann–Whitney *U* for two groups and Kruskal–Wallis for four groups) in addition to the follow-up ANOVAs, which confirmed significance of all significant ANOVA results. For the sake of brevity, these results are not reported here. With these follow-up analyses, we could describe, for example, that parental pre-knowledge led to parents asking more open-ended *wh*-questions, or that children with a higher inquiry attitude asked less closed questions.

RESULTS

Descriptions of Person Characteristics and Parent–Child Interaction

Person Characteristics

Mean values of person characteristics used in analyses are reported in **Table 1**, for all participants and for participants per pre-knowledge condition. A profile description of participating dyads can be found in Participants (see "Materials and Methods" section). On average, children agreed with enjoying science lessons ($M = 41.39$, $SD = 7.94$); they scored significantly higher than the international standard ($M = 32.8$, $SD = 9.5$; Fraser, 1981), $t(104) = 11.08$, $p < 0.001$. Note that in the international standard, children were older (12- and 13-year-olds) than in the current study. Interesting is that, although they enjoyed science lessons, on average ($M = 37.21$, $SD = 5.90$) children rated themselves as having a moderate inquiry attitude (30 = not being sure of having, 40 = agreeing with having an inquiry attitude).

ANOVA's (ratio variables) and chi-square analyses (nominal variables) were performed to investigate equal distribution of the person characteristics (the parents' gender and educational level, and the children's gender, age, working memory, enjoyment in science lessons, and inquiry attitude) across pre-knowledge conditions. No differences between pre-knowledge conditions were found.

Parent–Child Interaction

Summary values, and values per pre-knowledge condition, are reported in **Table 2**. On average, dyads inquired the inquiry task for 3.14 min. Dyads in the with pre-knowledge experimental condition ($M = 3.62$, $SD = 1.38$), played for 1 min longer than dyads in the without pre-knowledge control condition ($M = 2.63$, $SD = 1.52$), $t(103) = -3.483$; $p < 0.001$.

When manipulating, in most cases (73%), the child (M_C) or the parent (M_P) manipulated alone. There was a high correlation between not manipulating the black-box and content-related talk (sum of C1–C6) of parents, $r = 0.88$, $p < 0.001$, and children, $r = 0.67$, $p < 0.001$.

Parents, $M = 22.89$, $SD = 11.52$, contributed more than children ($M = 14.39$, $SD = 8.20$), to content-related talk during inquiry, $t(104) = 8.967$, $p < 0.001$. Parents made all kinds of content-related comments, but most often gave directions to the child (C5-P), "okay, so we have to find out how the ropes are running." Notably, they formulated almost no hypotheses (C6-P). Parents with pre-knowledge significantly contributed more to content-related talk ($M = 26.41$, $SD = 12.06$),

TABLE 1 | Factors and covariates, describing person characteristics of parent and child.

				Parental pre-knowledge	
			Total	Without	With
Parent					
Gender	Male	*N* (%)	50 (48%)	22 (43%)	28 (52%)
	Female	*N* (%)	55 (52%)	29 (57%)	26 (48%)
Dyad	Father–Son	*N* (%)	26 (25%)	12 (24%)	14 (26%)
	Father–Daughter	*N* (%)	24 (23%)	10 (20%)	14 (26%)
	Mother–Son	*N* (%)	25 (24%)	13 (25%)	12 (22%)
	Mother–Daughter	*N* (%)	30 (29%)	16 (31%)	14 (26%)
Educational level	Up to Bachelor (L)	*N* (%)	38 (36%)	17 (33%)	21 (39%)
	Bachelor (B)	*N* (%)	47 (45%)	22 (43%)	25 (46%)
	Graduate (G)	*N* (%)	20 (19%)	12 (24%)	8 (15%)
Child					
Gender	Male	*N* (%)	51 (49%)	25 (49%)	26 (48%)
	Female	*N* (%)	54 (51%)	26 (51%)	28 (52%)
Age		*M* (*SD*)	9.96 (1.38)	9.95 (1.36)	9.97 (1.42)
Working memory		*M* (*SD*)	4.49 (0.89)	4.50 (0.87)	4.49 (0.92)
Enjoyment in science lessons		*M* (*SD*)	41.39 (7.94)	41.84 (8.05)	40.96 (7.89)
Attitude to inquiry		*M* (*SD*)	37.21 (5.90)	37.39 (7.10)	37.04 (4.56)

Total number (N) and percentages (%) of the parents' and children's gender, the distribution of gender over the various parent–child dyads and the parents' educational level, and average values (M) and standard deviations (SD) of the children's age, working memory, enjoyment in science lessons and attitude to inquiry. Total = all participating parent–child dyads (N = 105). Without = parent–child dyads without parental pre-knowledge, control condition (N = 51). With = parent–child dyads with parental pre-knowledge, experimental condition (N = 54).

than parents without pre-knowledge (M = 19.16, SD = 9.71), $t(103) = 3.38$, $p < 0.001$.

Children most often interpreted results (C4-C), "I think these are two different ropes," described observations (C3-C), "all ropes have different colors," and made non-content-related comments (C7-C), such as expressing emotions "this is really difficult!". Children asked relatively few questions (C1-C, C2-C) and, similar to the parents, formulated almost no hypotheses (C6-C). Children of parents with pre-knowledge significantly contributed more to content-related talk (M = 16.33, SD = 7.83) than children of parents without pre-knowledge (M = 12.33, SD = 8.16), $t(103) = 2.56$, $p < 0.05$.

To give an impression of the conversations, two examples are displayed in **Table 3**. Hypotheses do occur (see example 1; "Yes, let's try that. So if I pull these two, then those two go over there."). However, sometimes participants observe the effect of a manipulation just before they fully express the hypothesis, and therefore the expression does not count as hypotheses (see in example 2; "And if we pull this, yes the yellow one goes more smoothly").

The Impact of Parental Pre-knowledge and Person Characteristics on Parent–Child Interaction

Overall Results

Manipulations

To find out if pre-knowledge and person characteristics affected parent and child's manipulations during inquiry, a MANCOVA was performed with the five manipulation categories (M_0, M_C, M_P, $M_{P/C}$, and $M_{P\&C}$) as outcome variables, with parental pre-knowledge, gender, and educational level, and children's gender as factors, and with children's age, working memory, enjoyment in science lessons and inquiry attitude as co-variates. Also, three two-way factor interactions were included (condition with respectively gender parent, educational level parent and gender child). Results showed significant main effects of parental pre-knowledge, $T = 0.35$, $F(5,87) = 6.11$, $p < 0.001$, $n^2 = 0.26$, and children's gender, $T = 0.14$, $F(5,87) = 2.36$, $p = 0.05$, $n^2 = 0.12$, on manipulations. In addition, there was an interaction effect of parental pre-knowledge and parent's gender on manipulations, $T = 0.14$, $F(5,87) = 2.38$, $p = 0.05$, $n^2 = 0.12$.

Parent talk

To find out if parental pre-knowledge and parent characteristics affected parent talk, a MANOVA was performed with six conversation categories (C1-P to C6-P) as outcome variables and with parental pre-knowledge, gender and educational level as factors. Also, the two two-way factor interactions were considered (condition with respectively parent's gender and educational level). Results showed a significant main effect of parental pre-knowledge [$T = 0.61$, $F(6,92) = 9.36$, $p < 0.001$, $n^2 = 0.38$] on parent talk. No other significant effects were observed.

Child talk

To find out if parental pre-knowledge and child characteristics affected child talk, a MANCOVA was performed with six conversation categories (C1-C to C6-C) as outcome variables, with parental pre-knowledge and children's gender as factors, and with children's age, working memory, enjoyment in science

TABLE 2 | Outcome variables, describing parent–child interaction during an inquiry activity in the museum.

				Pre-knowledge		Inquiry Attitude		Gender Child	
Parent–child Interaction		**Range *(Min–Max)***	**Total *M (SD)***	**Without *M (SD)***	**With *M (SD)***	**Low *M (SD)***	**High *M (SD)***	**Boys *M (SD)***	**Girls *M (SD)***
Manipulations									
No manipulations	M_0	5–55	24.87 (13.45)	**18.80** (11.29)	**30.59** (12.89)			22.76 (13.96)	26.85 (12.76)
Child manipulates alone	M_C	0–30	10.53 (6.71)	**8.67** (6.59)	**12.30** (6.40)			9.39 (5.95)	11.61 (7.25)
Parent manipulates alone	M_P	0–22	5.87 (5.19)	6.63 (5.29)	5.15 (5.04)			**5.02** (4.38)	**6.67** (5.78)
Parent and child in parallel	$M_{P/C}$	0–14	3.23 (3.21)	3.96 (3.39)	2.54 (2.89)			3.33 (3.12)	3.13 (3.32)
Parent and child together	$M_{P\&C}$	0–16	2.89 (3.72)	2.22 (2.77)	3.52 (4.36)			**3.55** (3.89)	**2.26** (3.48)
Parent Talk									
Asking *wh*-questions	C1-P	0–15	3.96 (3.39)	**2.10** (2.18)	**5.72** (3.40)				
Asking closed questions	C2-P	0–19	5.35 (4.12)	**3.86** (2.94)	**6.76** (4.58)				
Describing evidence	C3-P	0–11	3.16 (2.65)	2.96 (2.49)	3.35 (2.80)				
Interpreting results	C4-P	0–19	3.65 (3.25)	**4.33** (3.39)	**3.00** (3.00)				
Giving directions	C5-P	0–24	6.69 (4.02)	5.82 (3.60)	7.50 (4.25)				
Formulating hypotheses	C6-P	0–2	0.08 (0.30)	0.08 (0.27)	0.07 (0.33)				
Child Talk									
Asking *wh*-questions	C1-C	0–5	0.33 (0.76)	0.41 (0.90)	0.26 (0.59)	**0.18** (0.48)	**0.50** (0.95)		
Asking closed questions	C2-C	0–5	0.75 (1.10)	0.57 (0.70)	0.93 (1.36)	**0.98** (1.13)	**0.50** (0.74)		
Describing evidence	C3-C	0–19	4.76 (4.00)	**4.00** (4.05)	**5.48** (3.86)	4.16 (3.56)	5.42 (4.38)		
Interpreting results	C4-C	0–17	6.48 (3.83)	**5.25** (3.49)	**7.63** (3.80)	6.69 (4.37)	6.24 (3.15)		
Giving directions	C5-C	0–9	1.92 (2.09)	2.00 (2.30)	1.85 (1.90)	1.80 (1.98)	2.06 (2.23)		
Formulating hypotheses	C6-C	0–3	0.14 (0.49)	0.10 (0.36)	0.19 (0.59)	0.18 (0.61)	0.10 (0.30)		

Mean values (M) and standard deviations (SD), of three categories of outcome variables (manipulations, parent talk, and child talk), are presented for condition (parental pre-knowledge) and person characteristics (attitude to inquiry and gender child) with a main effect on parent–child interaction. Column "Total" contains range and mean values of all participating parent–child dyads (N = 105). Column "Without" and "With" contains values of parent–child dyads in the without parental pre-knowledge control condition (N = 51) and with parental pre-knowledge experimental condition (N = 54), respectively. Column "Low" and "High" contains values of parent–child dyads with children having a low (N = 55) and high (N = 50) attitude to inquiry, respectively. Column "Boys" and "Girls" contains values of parent–boy (N = 51) and parent–girl (N = 54) dyads, respectively. In bold: outcome variable with a significant difference between pre-knowledge conditions (without and with), inquiry attitude groups (low and high), and the child's gender (boys and girls) respectively.

lessons and inquiry attitude as co-variates. Also, the two-way factor interaction was taken into account (condition with gender child). Results showed significant main effects of parental pre-knowledge [$T = 0.18$, $F(6,92) = 2.76$, $p = 0.02$, $n^2 = 0.15$], and children's inquiry attitude [$T = 0.19$, $F(6,92) = 2.89$, $p = 0.01$, $n^2 = 0.16$] on child talk. No significant interaction effects on child talk were observed.

RQ-1

RQ-1: How does parental pre-knowledge affect parent–child interaction during an inquiry activity in the museum?

Manipulations

To find out how pre-knowledge affected parents' and children's manipulations during inquiry, the five separate univariate ANOVA's on the outcome variables are reported. Higher amounts of no manipulations (M_0), $F(1,91) = 15.63$, $p < 0.001$, $n^2 = 0.15$, and child manipulates alone (M_C), $F(1,91) = 9.00$, $p = 0.003$, $n^2 = 0.09$, were observed for parent–child dyads with parental pre-knowledge (see **Table 2**).

Parent talk

To find out how pre-knowledge affected parent talk, the six separate univariate ANOVAs on the outcome variables are reported. Higher amounts of asking open-ended *wh*-questions (C1-P), $F(1,97) = 32.38$, $p < 0.001$, $n^2 = 0.25$, asking closed questions (C2-P), $F(1,97) = 14.98$, $p < 0.001$, $n^2 = 0.13$, and lower amount of interpreting results (C4-P), $F(1,97) = 5.09$, $p = 0.03$, $n^2 = 0.05$, were observed for parents with parental pre-knowledge (see **Table 2**).

Child talk

To find out how pre-knowledge affected child talk, the six separate univariate ANOVAs on the outcome variables are reported. Higher amounts of describing evidence (C3-C), $F(1,97) = 4.43$, $p = 0.04$, $n^2 = 0.04$, and interpreting results (C4-C), $F(1,97) = 10.90$, $p = 0.001$, $n^2 = 0.10$, were observed for children of dyads with parental pre-knowledge, compared to children of dyads without parental pre-knowledge.

RQ-2

RQ-2: How do person characteristics affect parent–child interaction during an inquiry activity and the possible relation of parental pre-knowledge and parent–child interaction?

Manipulations

The follow-up univariate ANOVA's revealed that, compared to parent–son dyads, in parent–daughter dyads, parents more often manipulated alone (MP), $F(1,91) = 4.20$, $p = 0.04$, $n^2 = 0.04$, and parent–daughter dyads less often manipulated together ($M_{P\&C}$), $F(1,91) = 3.96$, $p = 0.05$, $n^2 = 0.04$. The follow-up univariate ANOVAs also revealed that only for father–child

TABLE 3 | Two examples of parent–child conversations during inquiry at a black-box.

Example 1			
Father:	Look, that one pulls, yellow pulls, ah!	C3-P	$M_{P/C}$
Daughter:	Now, pull that one.	C5-C	M_C
Father:	Look at that, this one goes to the middle.	C3-P	M_C
Daughter:	I know this one.	C7-C	$M_{P/C}$
Daughter:	This one goes, those two.	C3-C	M_P
Father:	Yes, and this one is.	C2-P	M_P
Daughter:	Also connected.	C4-C	M_P
Daughter:	Ah, I know.	C7-C	M_0
Father:	You already know?	C2-P	M_0
Daughter:	Ahu.	C7-C	M_0
Father:	What do you think? Well, tell me then.	C1-P	M_0
Daughter:	These two are together, and so are those, and they form a knot.	C4-C	M_C
Father:	Yes, let's try that. So if I pull these two, then those two go over there.	C6-P	M_P
Father:	Yes, and now you pull those two. Yes, and now we pull only one.	C5-P	$M_{P/C}$
Father:	Yes, I think you are right.	C7-P	M_0
Example 2			
Father:	You pull those and then those two go in.	C3-P	M_C
Son:	So uhm. then this one is underneath that one, I guess.	C4-C	M_0
Son:	Pull, for example, the red one.	C5-C	M_C
Son:	Then these two go. And if we pull this, yes the yellow one goes more smoothly.	C3-C	$M_{P\&C}$
Son:	So red goes with yellow.	C4-C	M_0
Father:	Now pull the yellow one.	C5-P	M_C
Father:	Then the blue one goes in.	C3-P	M_0
Father:	Stop, stop, otherwise you'll pull it all the way in.	C5-P	M_0
Son:	And with this one, the yellow one goes.	C3-C	$M_{P/C}$

Example 1 is a full conversation of father–daughter (without pre-knowledge; 8-year-old). Example 2 is a conversation fragment of father–son (without pre-knowledge; 9-year-old). C1, asking open-ended wh-questions; C2, asking closed questions; C3, describing evidence; C4, interpreting results; C5, giving directions; C6, formulating hypotheses; C7, other comments. P, Parent; C, Child. M_0, no manipulations; M_C, child manipulates alone; M_P, parent manipulates alone; $M_{P/C}$, parent and child in parallel; $M_{P\&C}$, parent and child together.

dyads, compared to dyads without parental pre-knowledge, children of dyads with parental pre-knowledge more often manipulated alone (M_C), $F(1,91) = 8.30$, $p = 0.005$, $n^2 = 0.08$ (see also **Figure 2**).

Child talk

To investigate how children's inquiry attitude affects child talk, the six separate univariate ANOVAs on the outcome variables are reported. Children with higher self-reported inquiry attitude more often asked open-ended *wh*-questions (C1-C), $F(1,97) = 6.10$, $p = 0.02$, $n^2 = 0.06$, and less often asked closed questions (C2-C), $F(1,97) = 3.96$, $p = 0.05$, $n^2 = 0.04$, compared to children with lower self-reported inquiry attitude (see **Table 2**).

Impact of Parental Pre-knowledge on Solution Accuracy

Parents' and children's solution accuracy, measured by classifying their drawings in one of four drawing categories, is presented in **Table 4**. Parental pre-knowledge resulted in higher accuracy for children ($\chi^2 = 12.88$, $p_{bootstrap} = 0.003$) and parents ($\chi^2 = 53.43$, $p_{bootstrap} < 0.001$). The most striking difference is that none of the children (0%) in the control condition were able to solve the black-box problem correctly, compared to 15% in the with parental pre-knowledge condition. Looking at fathers and mothers separately, it appears that in the pre-knowledge condition, fathers' solutions were as good as mothers' solutions ($\chi^2 = 1.443$, $p_{bootstrap} = 0.70$). However, in the control condition, fathers gave better solutions than mothers ($\chi^2 = 1.443$, $p_{bootstrap} = 0.04$).

DISCUSSION

The current study's main research question concerned the effect of parental pre-knowledge on parent–child interaction during inquiry. We used an experimental design to study this question, and additionally focused on the effects of parents' and children's person characteristics on parent–child interactions. Parent–child interactions were assessed by coding non-verbal (manipulations) and verbal behaviors (conversations), and learning was measured as solution accuracy after the parent and child's inquiry activity. To allow for manipulation of parental pre-knowledge, a black-box was used as the object of inquiry: a closed box where four rope ends stuck out. By inquiring, that is, pulling the ropes, dyads could largely figure out how the ropes were entangled inside the box. Parents in the pre-knowledge condition were shown the entangled ropes in the inside of the box prior to the inquiry activity, without children knowing.

The Effect of Parental Pre-knowledge on Parent–Child Interaction

Parental pre-knowledge led to differences in parent–child interaction with regard to inquiry time, manipulations, conversations, and learning. Parent and child in the parental pre-knowledge condition, inquired the box for a substantially longer period of time. Below, differences in manipulations, conversations, and learning will be discussed.

Manipulations

Children of parents with parental pre-knowledge more often manipulated the ropes on their own than children of parents without pre-knowledge. To the best of our knowledge, research that investigated the relation between parental knowledge and parent–child interaction during inquiry in the museum did not report on children's individual contribution to manipulations (Palmquist and Crowley, 2007; Benjamin et al., 2010; Eberbach and Crowley, 2017).

A second finding was that parents and children in the parental pre-knowledge condition more often did not manipulate the ropes. It seems that these dyads focused more on verbal

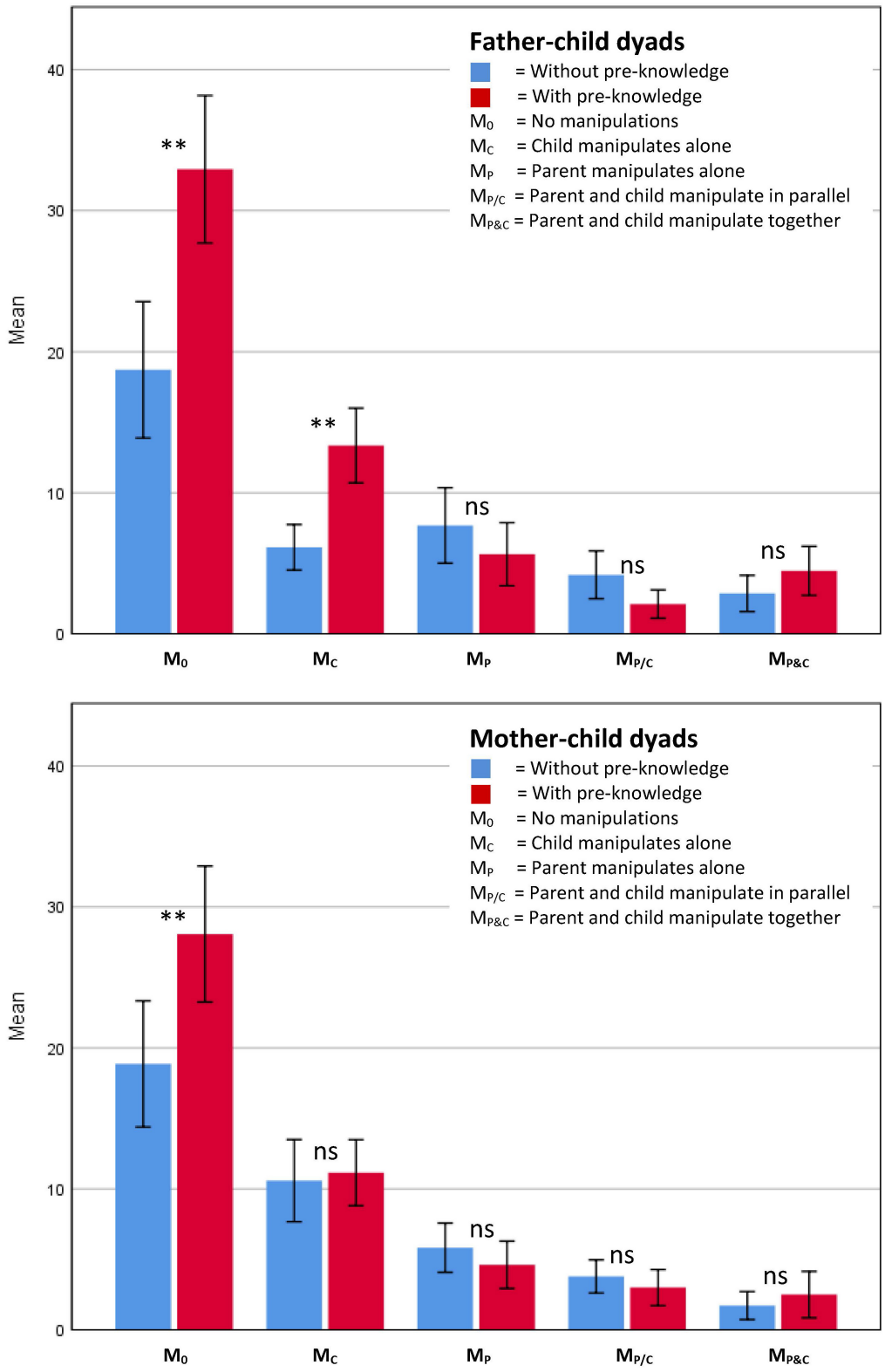


FIGURE 2 | Parental pre-knowledge and parental gender interaction on manipulations. **(A)** Father–child dyads (N = 50). **(B)** Mother–child dyads (N = 55). Mean values of the five manipulation categories (M_0, M_C, M_P, $M_{P/C}$, and $M_{P\&C}$) for the two pre-knowledge conditions (blue = Control condition without parental pre-knowledge, red = Experimental condition with parental pre-knowledge). Error bars = 95% CI. Significances for differences between pre-knowledge conditions are depicted: $^{ns}p > 0.05$; $^{**}p \leq 0.01$.

TABLE 4 | Parents' and children's solution accuracy of the inquiry activity.

		Without pre-knowledge		With pre-knowledge	
		Child	**Parent**	**Child**	**Parent**
K1 (incorrect)	*N* (%)	7 (14)	5 (10)	8 (15)	1 (2)
K2	*N* (%)	10 (20)	2 (4)	17 (31)	5 (9)
K3	*N* (%)	34 (67)	44 (86)	21 (39)	14 (26)
K4 (correct)	*N* (%)	0 (0)	0 (0)	8 (15)	34 (63)

Total number (N) and percentages (%) of parents and children per drawing category (K1, K2, K3, and K4). Column "Without" and "With" are parents (Parent) and children (Child) in the without (N = 51) and with (N = 54) parental pre-knowledge condition, respectively. K1, incorrect ropes and incorrect connection; K2, correct ropes but incorrect, fixed connection; K3, correct ropes but incorrect, loose connection; and K4, correct ropes and correct loose connection.

exchange than dyads in the condition without parental pre-knowledge. On average, these dyads indeed talked more, during which they did not manipulate. These dyads also inquired for a substantially longer period of time longer (i.e., on average 1 min more) compared to dyads in the without pre-knowledge condition. In line with these findings, Benjamin et al. (2010) report a positive causal effect of parents' elaborative talk (open-ended *wh*-questions and associations) on dyads' time spent in the exhibition.

Conversations

During the conversations, parents with pre-knowledge less often interpreted results themselves compared to parents without pre-knowledge. Instead, parents with pre-knowledge more often asked questions, both open-ended *wh*-questions and closed questions. That is, results indicate that parents with pre-knowledge behaved more as if they were in the role of being a supporter of the child's learning process by asking questions, instead of being a learner alongside the child. Especially asking open-ended *wh*-questions is considered to be an important strategy for supporting children's problem-solving and knowledge seeking behavior in informal (Boland et al., 2003; Eberbach and Crowley, 2017) and formal (Smith and Reiser, 2005) learning context. Children with parents who did have pre-knowledge more often described evidence and interpreted results. Several museum studies demonstrate positive effects of parents' *wh*-questions on children inquiry and learning (Crowley et al., 2001a; Benjamin et al., 2010; Willard et al., 2019).

Our finding that parental pre-knowledge enhanced asking open-ended *wh*-questions was not evidenced by Benjamin et al. (2010). An explanation for this discrepancy in findings could be that in Benjamin et al.'s (2010) study both parent and child received information, while in the current study, only the parents did, without children knowing. Our results are in line with those of Palmquist and Crowley (2007) who report that parents of novice children talked more and in a more supportive way with their children than parents with expert children. The same relationship is supported by the results of Eberbach and Crowley (2017) who report a correlation of parental pre-knowledge of pollination and the amount of open-ended *wh*-questions asked by parents during a botanical garden visit.

Children's Solution Accuracy

The finding that some of the children in the pre-knowledge condition solved the inquiry problem, compared to none of the children in the without pre-knowledge condition, indicated that parental pre-knowledge facilitated children's learning. One explanation is that pre-knowledge facilitated parents in scaffolding their children's learning through inquiry. This explanation is supported by higher amounts of open-ended *wh*-questions that parents in the pre-knowledge condition asked their children. The explanation that parents with pre-knowledge told the inquiry problem's solution directly to their child is less likely since parents with pre-knowledge less often interpreted results, which in the current study entails giving explanations. A second finding was that, in the control condition, not only the children but also the parents were often unable to solve the inquiry problem in all detail. However, partially correct answers were given by a substantial group of parents and children. That is, by inquiry they could solve important aspects of the task. Nevertheless, this suggests that the task was difficult to solve fully by inquiry alone, that is, without guidance by a parent with pre-knowledge. As the analysis of the conversation shows, the contribution of the parent with pre-knowledge was not explaining the solution, but scaffolding the child in discovering the solution. These results are in line with insights from education effect studies into open-ended inquiry learning (Alfieri et al., 2011) and studies into guided inquiry learning (Van de Pol et al., 2010).

The Impact of Person Characteristics on Parent–Child Interaction

This study did show ample evidence for the impact of child characteristics on parent–child interaction. Parent characteristics were only in an interaction between gender and parental pre-knowledge related to parent and child's manipulations.

Child Characteristics

The child's gender was found to be related to how parent and child cooperated during manipulations. Compared to parent–son dyads, in parent–daughter dyads, parents more often manipulated alone and parent and child less often manipulated together during the inquiry activity. Informal science education literature reports on differences in how parents interact with boys and girls while engaging in inquiry activities (Crowley et al., 2001b; Siegel et al., 2007; Luce et al., 2013). As observed in the current study, Siegel et al. (2007) found that, at home and with a closed-ended task, parents behaved more collaboratively with boys than with girls. However, in the same study, in a museum context and with more open-ended activities, parents collaborate more with girls than with boys. Apparently, correlations between the child's gender and parent–child interaction are highly context and task specific. An explanation for context-related differences in parental behavior with girls and boys, is parents' gender-biased beliefs about children's abilities and interests (Sigel and McGillicuddy-De Lisi, 2002; Siegel et al., 2007). In future research, to better understand how the child's gender relates to parent–child interaction during inquiry, it could be informative

to examine the impact of parents' beliefs about their child's science achievement and science interest.

Another explanation for the observed differences in cooperation in the current study is that, instead of parents, children acted differently. Whether cooperation is achieved by the child, or is a reaction of the child to the parents' behavior (or the parent to the child's behavior) is difficult to disentangle (Tenenbaum and Leaper, 2003). Anecdotally Siegel et al. (2007) observed that boys, more than girls, were taking the lead in performing experiments. In science education literature, the relation between the child's gender and science achievement has been studied with contradicting results: some studies showed significant differences between boys and girls, while others did not (Sungur and Tekkaya, 2003).

Children's self-reported inquiry attitude was related to differences in child talk. Children with higher inquiry attitude (e.g., children who prefer to solve a problem by inquiring instead of by being told) more often asked open-ended *wh*-questions, and less often asked closed questions, compared to children with lower inquiry attitude. This relation could be considered as a validation of the TOSRA Inquiry Attitude subscale (Fraser, 1981) in the museum context. Attitude toward science is recognized as an important educational outcome as it relates to lifelong learning (Organisation for Economic Co-operation, and Development [OECD], 2009; Ainley and Ainley, 2011; Sachisthal et al., 2018).

Parent Characteristics

Against our expectations, no main effects of parent characteristics on parent–child interaction were observed, besides the experimentally acquired parental pre-knowledge. Museum literature shows that dyads' time spent in the exhibition was found to be correlated to parents' educational level (Szechter and Carey, 2009) and parents' attitude toward science (Tare et al., 2011). Moreover, parent–child interaction was found to be correlated to the parents' gender (Brown, 1995; Benjamin et al., 2010; Nadelson, 2013; Van Schijndel and Raijmakers, 2016). For example, compared to mothers, fathers tended to be more active when accompanying their child's exploration of hands-on exhibits (Brown, 1995). Compared to fathers, mothers gave more causal explanations (Van Schijndel and Raijmakers, 2016). However, we did observe an interaction effect of parents' gender and parental pre-knowledge on manipulations. The interaction effect of parental pre-knowledge and parent gender on manipulations (see also *Section "Conversations"*) revealed that children of father–child dyads, not of mother-child dyads, manipulated more frequently alone in case the parent had pre-knowledge. Relating this result to the accuracy of solutions found by the parents, it seems that fathers without pre-knowledge (relative to fathers with pre-knowledge) were giving children less time to explore alone, because they were finding out the correct solution. Mothers without pre-knowledge did not limit the time of their children to explore alone, resulting in worse solutions than the fathers without pre-knowledge.

Limitations of the Study

In studying the causal effect of pre-knowledge on parent–child interaction, we purposefully choose a decontextualized inquiry activity (Clough, 2006), to ensure that none of the participants had pre-knowledge about the inquiry activity (i.e., the black-box). Most inquiry exhibits in science museums, however, are contextualized inquiry exhibits based on natural phenomena. One could question therefore, the external validity of the task for science museums practices. After all, black-box activities challenge the acquisition and use of domain-specific knowledge less. However, black-box activities provide similar challenges as encountered when inquiring natural phenomena; they stimulate the use of domain-general strategies as asking questions, experimenting, observing, and interpreting evidence (Lederman and Abd-El-Khalick, 1998). Museum exhibits that foster these inquiry skills, such as APE (Active Prolonged Engagement) exhibits, do not depend strongly on pre-knowledge, in contrast to counterintuitive exhibits (Gutwill, 2008). Our data showed that parents with pre-knowledge did not share their pre-knowledge directly with the child; parents even explained less to their child. Whether this effect of parental pre-knowledge on parent–child interaction will be the same for phenomenon-based inquiry exhibits, we cannot tell. We imagine that having information about "how to solve a specific problem" (black-box) is different from having information about "how a phenomenon works." In the latter case, the information has a value that goes beyond the situation of the inquiry activity. This "eternity" value could motivate parents to act differently, for example, by sharing this information with the child by interpreting results in addition to asking open-ended *wh*-questions. Also, knowledge about a real-world phenomenon could make parents more interested in the inquiry activity (Tare et al., 2011; Ainley, 2017; Sachisthal et al., 2018) and this could raise new questions that the parents might want to inquire. In future research, it could be informative to further investigate the effect of parental pre-knowledge on parent–child interaction for a series of phenomena, which differ in the extent that they are familiar to the parents.

A second limitation is the generalizability of the outcomes to natural settings. Our test design does not reflect the natural situation of families inquiring at home or during a free-choice science museum visit. The research setting and the presence of the video camera could have motivated participants extrinsically to inquire longer and with more attention, resulting in a longer inquiry time and more interaction between parent and child (Pattison and Shagott, 2015). However, these possible motivating circumstances were the same for both pre-knowledge conditions.

A third limitation is the generalizability of the study outcomes to non-museum settings. Participants were recruited from the museum population and on average were higher educated than the average Dutch. This could have impacted parent–child interaction; however, our data did not show an effect of parental education level on parent–child interaction.

Implications for Museum Practice

Insights from research about how person characteristics impact family inquiry in the museum are of value for museum practice, for example, when designing for specific audiences (e.g., Dritsas et al., 1998; Dancstep and Sindorf, 2018) or specific learning experiences (e.g., Humphrey and Gutwill, 2005; Gutwill, 2008; Povis and Crowley, 2015). The findings of the current study could

be of interest for museum professionals in making informed choices in exhibition design in relation to desired objectives. The current study with the black-box as activity seems especially relevant for exhibits that foster inquiry skills. The current study suggests, for example, that if aiming to support parents in their role of scaffolding the child's learning, then it may be helpful to opt for phenomena that parents are more familiar with or to provide parents with information about the specific phenomenon. Our results showed that parents with domain-specific pre-knowledge more often scaffolded their child by asking open-ended *wh*-questions. It seems essential that only the parent has this domain-specific pre-knowledge, not the child (cf. Benjamin et al., 2010). Museum research has shown that parental scaffolding behavior can be encouraged by pre-visit instruction (e.g., domain-general process knowledge), for example, an inquiry training (Gutwill and Allen, 2010), an instruction to use elaborative speech (Benjamin et al., 2010), an instruction to explain (Willard et al., 2019), and an instructional video about coaching techniques (Van Schijndel et al., 2010). An interesting aspect of the current study is that, by offering parents a domain-specific knowledge edge (in contrast to scaffolding or inquiry instruction), they spontaneously showed scaffolding behavior by asking more open-ended *wh*-questions without being trained or instructed to do so.

CONCLUSION

Children's science learning is for an important part dependent on how families observe, discuss, and explore science and technology (National Research Council, 2009; Haden, 2010). The current study investigated how person characteristics relate to families' learning from inquiry activities and demonstrates that parental pre-knowledge affects the way parents interact and explore with their child. Compared to parents without pre-knowledge, parents with pre-knowledge inquired longer, posed more open-ended *wh*-questions and closed questions, and less often interpreted results. The children of parents with pre-knowledge more often manipulated alone, more often described evidence, more often interpreted results, and gave better solutions. In addition, the study demonstrates that child characteristics affected parent–child interaction during inquiry. Boys more often than girls cooperated with their parents, girls more often than boys manipulated alone, and children with a self-reported higher inquiry attitude asked more open-ended *wh*-questions than children with a lower inquiry attitude. By offering parents a knowledge edge, they spontaneously showed scaffolding behavior by asking more open-ended *wh*-questions and they left the interpretation of inquiry results to their children without being trained or instructed to do so. The current study gives an insight into the potential effect of pre-knowledge on parent–child interaction during an inquiry activity and shows that having pre-knowledge can facilitate parents scaffolding behavior and can lead to a different learning situation for both child and parent.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Review Board of the Institute of Education and Child Studies, University of Leiden. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

RF and MR conceived the ideas and analyzed the data. MR designed the study and supervised the data collection. RF wrote the manuscript. TV and MR revised the manuscript. All authors have approved the final manuscript.

ACKNOWLEDGMENTS

We thank the parents and children who participated in this research. Also, we thank Kim Haslinghuijs and Iris van de Nes for their contribution to this study as part of their master theses.

REFERENCES

Ainley, M. (2017). "Interest: knowns, unknowns, and basic processes," in *The Science of Interest*, eds P. O'Keefe, and J. Harackiewicz (Cham: Springer), 3–24. doi: 10.1007/978-3-319-55509-6_1

Ainley, M., and Ainley, J. (2011). A cultural perspective on the structure of student interest in science. *Int. J. Sci. Educ.* 33, 51–71. doi: 10.1080/09500693.2010.518640

Alfieri, L., Brooks, P. J., Aldrich, N. J., and Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *J. Educ. Psychol.* 103, 1–18. doi: 10.1037/a0021017

Allen, S. (2002). "Looking for learning in visitor talk: a methodological exploration," in *Learning Conversations in Museums*, eds G. Leinhardt, K. Crowley, and K. Knutson (Mahwah, NJ: Lawrence Erlbaum), 259–303.

Allen, S. (2004). Designs for learning: studying science museum exhibits that do more than entertain. *Sci. Educ.* 88, S17–S33.

Ash, D. (2004). How families use questions at dioramas: ideas for exhibit design. *Curator* 47, 84–100. doi: 10.1111/j.2151-6952.2004.tb00367.x

Benjamin, N., Haden, C. A., and Wilkerson, E. (2010). Enhancing building, conversation, and learning through caregiver–child interactions in a children's museum. *Dev. Psychol.* 46, 502–515. doi: 10.1037/a0017822

Blanca, M. J., Alarcón, R., Arnau, J., Bono, R., and Bendayan, R. (2018). Effect of variance ratio on ANOVA robustness: might 1.5 be the limit? *Behav. Res. Methods* 50, 937–962. doi: 10.3758/s13428-017-0918-2

Boland, A. M., Haden, C. A., and Ornstein, P. A. (2003). Boosting children's memory by training mothers in the use of an elaborative conversational style as an event unfolds. *J. Cogn. Dev.* 4, 39–65. doi: 10.1080/15248372.2003.9669682

Brigham, F. J., Scruggs, T. E., and Mastropieri, M. A. (2011). Science education and students with learning disabilities. *Learn. Disabil. Res. Pract.* 26, 223–232. doi: 10.1111/j.1540-5826.2011.00343.x

Brown, C. (1995). Making the most of family visits: some observations of parents with children in a museum science centre. *Mus. Manag. Curatorship* 14, 65–71. doi: 10.1080/09647779509515428

Callanan, M. A., and Jipson, J. L. (2001). "Children's developing scientific literacy," in *Designing for Science: Implications from Everyday, Classroom, and Professional Settings*, eds K. Crowley, C. Shunn, and T. Okada (Mahwah, NJ: Lawrence Erlbaum), 19–43.

Callanan, M. A., and Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: causal thinking in everyday activity. *Cogn. Dev.* 7, 213–233. doi: 10.1016/0885-2014(92)90012-g

Callanan, M., and Valle, A. (2008). Co-constructing conceptual domains through family conversations and activities. *Psychol. Learn. Motiv.* 49, 147–165. doi: 10.1016/s0079-7421(08)00004-2

Chinn, C. A., and Malhotra, B. A. (2002). Children's responses to anomalous scientific data: how is conceptual change impeded? *J. Educ. Psychol.* 94, 327–343. doi: 10.1037/0022-0663.94.2.327

Clough, M. P. (2006). Learners' responses to the demands of conceptual change: considerations for effective nature of science instruction. *Sci. Educ.* 15, 463–494. doi: 10.1007/s11191-005-4846-7

Corsi, P. M. (1972). *Human Memory and the Medial Temporal Lobe Region of the Brain*. Doctoral dissertation, McGill University, Montreal.

Crowley, K., and Callanan, M. A. (1998). Describing and supporting collaborative scientific thinking in parent– child interactions. *J. Mus. Educ.* 23, 12–17. doi: 10.1080/10598650.1998.11510365

Crowley, K., and Jacobs, M. (2002). "Building islands of expertise," in *Learning Conversations in Museums*, eds G. Leinhardt, K. Crowley, and K. Knutson (Mahwah, NJ: Lawrence Erlbaum), 333–356.

Crowley, K., Callanan, M. A., Jipson, J. L., Galco, J., Topping, K., and Shrager, J. (2001a). Shared scientific thinking in everyday parent-child activity. *Sci. Educ.* 85, 712–732. doi: 10.1136/bmj.f2882

Crowley, K., Callanan, M. A., Tenenbaum, H. R., and Allen, E. (2001b). Parents explain more often to boys than to girls during shared scientific thinking. *Psychol. Sci.* 12, 258–261. doi: 10.1111/1467-9280.00347

Dancstep, T., and Sindorf, L. (2018). Creating a female-responsive design framework for STEM exhibits. *Curator* 61, 469–484. doi: 10.1111/cura.12268

Dovis, S., Van der Oord, S., Wiers, R. W., and Prins, P. J. (2012). Can motivation normalize working memory and task persistence in children with attention-deficit/hyperactivity disorder? The effects of money and computer-gaming. *J. Abnorm. Child Psychol.* 40, 669–681. doi: 10.1007/s10802-011-9601-8

Dritsas, J., Borun, M., and Johnson, J. I. (1998). *Family Learning in Museums: the PISEC Perspective. Philadelphia Camden Informal Science Education Collaborative (PISEC)*. Philadelphia, PA: The Franklin Institute.

Eberbach, C., and Crowley, K. (2017). From seeing to observing: how parents and children learn to see science in a botanical garden. *J. Learn. Sci.* 26, 608–642. doi: 10.1080/10508406.2017.1308867

Falk, J. H. (2009). *Identity and the Museum Visitor Experience*. Walnut Creek, CA: Left Coast Press.

Falk, J. H., and Dierking, L. D. (2000). *Learning from Museums: Visitor Experience and the Making of Meaning*. Walnut Creek, CA: Alta Mira Press.

Fender, J. G., and Crowley, K. (2007). How parent explanation changes what children learn from everyday scientific thinking. *J. Appl. Dev. Psychol.* 28, 189–210. doi: 10.1016/j.appdev.2007.02.007

Field, A. (2009). *Discovering Statistics Using SPSS*, 3rd Edn. Thousand Oaks, CA: Sage.

Fraser, B. J. (1981). *Tosra: Test of Science-Related Attitudes: Handbook*. Camberwell: Australian Council for Educational Research.

Gauvain, M. (2001). *The Social Context of Cognitive Development*. New York, NY: Guilford Press.

Geerdts, M. S., Van de Walle, G. A., and LoBue, V. (2015). Parent–child conversations about animals in informal learning environments. *Visit. Stud.* 18, 39–63. doi: 10.1080/10645578.2015.1016366

Gleason, M. E., and Schauble, L. (1999). Parents' assistance of their children's scientific reasoning. *Cogn. Instr.* 17, 343–378. doi: 10.1207/s1532690xci1704_1

Gutwill, J. P. (2008). Challenging a common assumption of hands-on exhibits: how counter-intuitive phenomena can undermine open-endedness. *J. Mus. Educ.* 33, 187–198. doi: 10.1080/10598650.2008.11510599

Gutwill, J. P., and Allen, S. (2010). Facilitating family group inquiry at science museum exhibits. *Sci. Educ.* 94, 710–742. doi: 10.1002/sce.20387

Haden, C. A. (2010). Talking about science in museums. *Child Dev. Perspect.* 4, 62–67. doi: 10.1111/j.1750-8606.2009.00119.x

Humphrey, T., and Gutwill, J. P. (eds). (2005). *Fostering Active Prolonged Engagement: The Art of Creating APE Exhibits*. Walnut Creek, CA: Left Coast Press.

Jipson, J. L., and Callanan, M. A. (2003). Mother–child conversation and children's understanding of biological and nonbiological changes in size. *Child Dev.* 74, 629–644. doi: 10.1111/1467-8624.7402020

Kirschner, P. A., Sweller, J., and Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educ. Psychol.* 41, 75–86. doi: 10.1016/j.jecp.2016.03.009

Kisiel, J., Rowe, S., Vartabedian, M. A., and Kopczak, C. (2012). Evidence for family engagement in scientific reasoning at interactive animal exhibits. *Sci. Educ.* 96, 1047–1070. doi: 10.1002/sce.21036

Klahr, D., and Dunbar, K. (1988). Dual search space during scientific reasoning. *Cogn. Sci.* 12, 1–48. doi: 10.1207/s15516709cog1201_1

Klahr, D., Fay, A. L., and Dunbar, K. (1993). Heuristics for scientific experimentation: a developmental study. *Cogn. Psychol.* 25, 111–146. doi: 10.1006/cogp.1993.1003

Knutson, K., and Crowley, K. (2010). "Connecting with art: how families talk about art in a museum setting," in *Instructional Explanations in the Disciplines*, eds M. K. Stein, and L. Kucan (Boston, MA: Springer), 189–206. doi: 10.1007/978-1-4419-0594-9_12

Kuhn, D., Amsel, E., and O'Loughlin, M. (1988). *The Development of Scientific Thinking Skills*. Orlando, FL: Academic Press.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.

Lederman, N. G., and Abd-El-Khalick, F. (1998). *The Nature of Science in Science Education: Rationales and Strategies*. Berlin: Springer.

Legare, C. H., Sobel, D. M., and Callanan, M. (2017). Causal learning is collaborative: examining explanation and exploration in social contexts. *Psychon. Bull. Rev.* 24, 1548–1554. doi: 10.3758/s13423-017-1351-3

Leinhardt, G., and Knutson, K. (2004). *Listening in on Museum Conversations*. Lanham, MD: Rowman Altamira.

Luce, M. R., Callanan, M. A., and Smilovic, S. (2013). Links between parents' epistemological stance and children's evidence talk. *Dev. Psychol.* 49, 454–461. doi: 10.1037/a0031249

MacWhinney, B., and Snow, C. (1990). The child language data exchange system: an update. *J. Child Lang.* 17, 457–472. doi: 10.1017/s0305000900013866

Marcus, M., Haden, C. A., and Uttal, D. H. (2018). Promoting children's learning and transfer across informal science, technology, engineering, and mathematics learning experiences. *J. Exp. Child Psychol.* 175, 80–95. doi: 10.1016/j.jecp.2018.06.003

Miller, S. (2014). Modeling the nature of science with the mystery tube. *Phys. Teach.* 52, 548–551. doi: 10.1119/1.4902200

Nadelson, L. S. (2013). Who is watching and who is playing: parental engagement with children at a hands-on science center. *J. Educ. Res.* 106, 478–484. doi: 10.1080/00220671.2013.833010

National Research Council, (2009). *Learning Science in Informal Environments: People, Places, and Pursuits*. Washington, DC: National Academies Press.

Organisation for Economic Co-operation, and Development [OECD] (2009). *Equally Prepared for Life? How 15-Year-Old Boys and Girls Perform in School*. Brussels: OECD.

Palmquist, S., and Crowley, K. (2007). From teachers to testers: how parents talk to novice and expert children in a natural history museum. *Sci. Educ.* 91, 783–804. doi: 10.1002/sce.20215

Pattison, S. A., and Dierking, L. D. (2013). Staff-mediated learning in museums: a social interaction perspective. *Visit. Stud.* 16, 117–143. doi: 10.1080/10645578.2013.767731

Pattison, S. A., and Shagott, T. (2015). Participant reactivity in museum research: the effect of cueing visitors at an interactive exhibit. *Visit. Stud.* 18, 214–232. doi: 10.1080/10645578.2015.1079103

Povis, K. T., and Crowley, K. (2015). Family learning in object-based museums: the role of joint attention. *Visit. Stud.* 18, 168–182. doi: 10.1080/10645578.2015.1079095

Ricco, R., and Rodriguez, P. (2006). The relation of personal epistemology to parenting style and goal orientation in college mothers. *Soc. Psychol. Educ.* 9, 159–178. doi: 10.1007/s11218-006-0001-9

Rosenshine, B. (2009). "The empirical support for direct instruction," in *Constructivist Instruction*, eds S. Tobias, and T. M. Duffy (Abingdon: Routledge), 201–220.

Sachisthal, M. S., Jansen, B. R., Peetsma, T. T., Dalege, J., van der Maas, H. L., and Raijmakers, M. E. (2018). Introducing a science interest network model to reveal country differences. *J. Educ. Psychol.* 111, 1063–1080. doi: 10.1037/edu0000327

Schauble, L. (1990). Belief revision in children: the role of prior knowledge and strategies for generating evidence. *J. Exp. Child Psychol.* 49, 31–57. doi: 10.1016/0022-0965(90)90048-d

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Dev. Psychol.* 32, 102–119. doi: 10.1037/0012-1649.32.1.102

Schauble, L., Gleason, M., Lehrer, R., Bartlett, K., Petrosino, A., Allen, A., et al. (2002). "Supporting science learning in museums," in *Learning Conversations in Museums*, eds G. Leinhardt, K. Crowley, and K. Knutson (Mahwah, NJ: Lawrence Erlbaum), 425–452.

Siegel, D. R., Esterly, J., Callanan, M. A., Wright, R., and Navarro, R. (2007). Conversations about science across activities in Mexican-descent families. *Int. J. Sci. Educ.* 29, 1447–1466. doi: 10.1080/09500690701494100

Sigel, I. E. (1998). *Socialization of Cognition: A Family Focus. Families, Risk, and Competence*. Mahwah, NJ: LEA, 289–307.

Sigel, I. E., and McGillicuddy-De Lisi, A. (2002). "Parent beliefs are cognitions: the dynamic belief systems model," in *Handbook of Parenting: Being and Becoming a Parent*, 2nd Edn, Vol. 3, ed. M. H. Bornstein (Mahwah, NJ: LEA), 485–508.

Smith, B. K., and Reiser, B. J. (2005). Explaining behavior through observational investigation and theory articulation. *J. Learn. Sci.* 14, 315–360. doi: 10.1207/s15327809jls1403_1

Sobel, D. M., and Jipson, J. L. (2015). *Cognitive Development in Museum Settings: Relating Research and Practice*. New York, NY: Routledge.

Sungur, S., and Tekkaya, C. (2003). Students' achievement in human circulatory system unit: the effect of reasoning ability and gender. *J. Sci. Educ. Technol.* 12, 59–64.

Szechter, L. E., and Carey, E. J. (2009). Gravitating toward science: parent–child interactions at a gravitational-wave observatory. *Sci. Educ.* 93, 846–858. doi: 10.1002/sce.20333

Tare, M., French, J., Frazier, B. N., Diamond, J., and Evans, E. M. (2011). Explanatory parent–child conversation predominates at an evolution exhibit. *Sci. Educ.* 95, 720–744. doi: 10.1002/sce.20433

Tenenbaum, H. R., and Hohenstein, J. M. (2016). Parent–child talk about the origins of living things. *J. Exp. Child Psychol.* 150, 314–329. doi: 10.1016/j.jecp.2016.06.007

Tenenbaum, H. R., and Leaper, C. (2003). Parent-child conversations about science: the socialization of gender inequities? *Dev. Psychol.* 39, 34–47. doi: 10.1037/0012-1649.39.1.34

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., et al. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: a review of literature. *J. Educ. Gifted* 27, 119–145. doi: 10.1177/016235320302700203

Trumbull, D. J., Bonney, R., and Grudens-Schuck, N. (2005). Developing materials to promote inquiry: lessons learned. *Sci. Educ.* 89, 879–900. doi: 10.1002/sce.20081

Van de Pol, J., Volman, M., and Beishuizen, J. (2010). Scaffolding in teacher–student interaction: a decade of research. *Educ. Psychol. Rev.* 22, 271–296. doi: 10.1007/s10648-010-9127-6

Van Schijndel, T. J., and Raijmakers, M. E. (2016). Parent explanation and preschoolers' exploratory behavior and learning in a shadow exhibition. *Sci. Educ.* 100, 153–178. doi: 10.1002/sce.21193

Van Schijndel, T. J., Franse, R. K., and Raijmakers, M. E. (2010). The exploratory behavior scale: assessing young visitors' hands-on behavior in science museums. *Sci. Educ.* 94, 794–809. doi: 10.1002/sce.20394

Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., and Deci, E. L. (2004). Motivating learning, performance, and persistence: the synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *J. Pers. Soc. Psychol.* 87, 246–260. doi: 10.1037/0022-3514.87.2.246

Walma van der Molen, J., Wiegerinck, S., and Rohaan, E. J. (2007). *Final Report VTB Attitude Monitor: The Development of an Attitude Instrument in the Field of Science and Technology for Pupils in Primary Education*. Amsterdam: JWVDM-Advies, training en onderzoek.

Watt, S. J., Therrien, W. J., and Kaldenberg, E. R. (2014). Meeting the diverse needs of students with EBD in inclusive science classrooms. *Beyond Behav.* 23, 14–19. doi: 10.1177/107429561402300203

Wechsler, D. (1958). *The Measurement and Appraisal of Adult Intelligence*, 4th Edn. Baltimore, MD: Williams & Wilkins Co.

Willard, A. K., Busch, J. T., Cullum, K. A., Letourneau, S. M., Sobel, D. M., Callanan, M., et al. (2019). Explain this, explore that: a study of parent–child interaction in a children's museum. *Child Dev.* 90, e598–e617. doi: 10.1111/cdev.13232

Wood, D., Bruner, J. S., and Ross, G. (1976). The role of tutoring in problem-solving. *J. Child Psychol. Psychiatry* 17, 89–100. doi: 10.1111/j.1469-7610.1976.tb00381.x

Zimmerman, C. (2005). *The Development of Scientific Reasoning Skills: What Psychologists Contribute to an Understanding of Elementary Science Learning. Final Draft of a Report to the National Research Council Committee on Science Learning Kindergarten through Eighth Grade*. Washington, DC: National Research Council.

Zweers, I., Huizinga, M., Denessen, E., and Raijmakers, M. E. J. (2019). *Inquiry-Based Learning For All: A Systematic Review of the Effects of Inquiry-Based Learning on Knowledge, Skills, Attitudes and Behavior of Students with Social-Emotional and Behavioral Difficulties in Primary and Secondary Education*. Available online at: https://doi.org/10.31219/osf.io/z45jt

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Franse, Van Schijndel and Raijmakers. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 10 June 2020
doi: 10.3389/fpsyg.2020.01016

Embedding Scientific Explanations Into Storybooks Impacts Children's Scientific Discourse and Learning

Kathryn A. Leech[1], Amanda S. Haber[2], Youmna Jalkh[2] and Kathleen H. Corriveau[2]*

[1] School of Education, The University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, [2] Wheelock College of Education and Applied Human Development, Boston University, Boston, MA, United States

Children's understanding of unobservable scientific entities largely depends on testimony from others, especially through parental explanations that highlight the mechanism underlying a scientific entity. Mechanistic explanations are particularly helpful in promoting children's conceptual understanding, yet they are relatively rare in parent–child conversations. The current study aimed to increase parent–child use of mechanistic conversation by modeling this language in a storybook about the mechanism of electrical circuits. We also examined whether an increase in mechanistic conversation was associated with science learning outcomes, measured at both the dyadic- and child-level. In the current study, parents and their 4- to 5-year-old children (N = 60) were randomly assigned to read a book containing mechanistic explanations (n = 32) or one containing non-mechanistic explanations (n = 28). After reading the book together, parent–child joint understanding of electricity's mechanism was tested by asking the dyad to assemble electrical components of a circuit toy so that a light would turn on. Finally, child science learning outcomes were examined by asking children to assemble a novel circuit toy and answer comprehension questions to gauge their understanding of electricity's mechanism. Results indicate that dyads who read storybooks containing mechanistic explanations were (1) more successful at completing the circuit (putting the pieces together to make the light turn on) and (2) used more mechanistic language than dyads assigned to the non-mechanistic condition. Children in the mechanistic condition also had better learning outcomes, but only if they engaged in more mechanistic discourse with their parent. We discuss these results using a social interactionist framework to highlight the role of input and interaction for learning. We also highlight how these results implicate everyday routines such as book reading in supporting children's scientific discourse and understanding.

Keywords: book reading, explanations, parent–child interaction, scientific discourse, social interaction

OPEN ACCESS

Edited by:
Amanda C. Brandone, Lehigh University, United States

Reviewed by:
Dorit Aram, Tel Aviv University, Israel
Jennifer Jipson, California Polytechnic State University, United States

****Correspondence:***
Kathryn A. Leech
leechk@unc.edu;
kathryn_leech@gse.harvard.edu

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *01 February 2020*
Accepted: *23 April 2020*
Published: *10 June 2020*

Citation:
Leech KA, Haber AS, Jalkh Y and Corriveau KH (2020) Embedding Scientific Explanations Into Storybooks Impacts Children's Scientific Discourse and Learning. Front. Psychol. 11:1016. doi: 10.3389/fpsyg.2020.01016

INTRODUCTION

Although children rely on their own exploration and experimentation to learn about everyday scientific phenomena, this investigation alone is not always sufficient for children's learning (Gelman et al., 2010; Harris and Corriveau, 2014; Legare, 2014). For example, when mechanisms that underlie a causal process are opaque or abstract, such as how electricity flows, children's

learning relies in part on information from others (Jipson et al., 2016; Harris et al., 2018). One type of information known to facilitate learning is adult explanations (e.g., Callanan and Oakes, 1992; Legare and Lombrozo, 2014; Lombrozo et al., 2018; Willard et al., 2019). However, as we describe in the text that follows, adults' explanations to preschool-aged children vary in their frequency as well as in features that impact children's learning (e.g., argument circularity, syntactic complexity, presence of causal mechanisms). Therefore, the primary aim of this study was to support adults in providing explanations that contain features that help children learn about abstract, unobservable scientific concepts. To do this, we designed a brief intervention in which features of scientific explanations were manipulated within a book-reading interaction between parents and their 4- to 5-year-old children. We investigated how such a manipulation impacted subsequent parent–child science interactions and children's independent scientific thinking.

Social Interactionist Theories of Learning

This study is motivated by social interactionist theories of development that state that children learn via input from and interaction with more knowledgeable others (Vygotsky, 1978; Bruner, 1983). Under this framework, interactions such as conversation between an adult and child support learning by helping the child organize knowledge and transfer such knowledge to novel situations. For example, talk about numbers relates to children's early mathematical understanding (e.g., Ramani et al., 2015), discussions that contain spatial language relate to children's spatial knowledge (e.g., Pruden et al., 2011), and references to emotions or mental states relate to children's socio-emotional development (e.g., Lagattuta and Wellman, 2002; Ziv et al., 2013). Here, we explore how parent–child conversations about science support children's scientific thinking and understanding. There are many features of science conversations that help children learn about scientific information, for instance, adults' questions that highlight important information, connections to previous experiences, and explanations that elucidate unobservable scientific mechanisms (Beals and Snow, 1994; Snow and Kurland, 1996; Crowley and Siegler, 1999; Beals, 2001; Callanan and Jipson, 2001; Crowley et al., 2001a,b; Tenenbaum et al., 2005; Haden, 2010). The current article focuses specifically on the relation between explanations in parent–child conversation and children's subsequent science learning.

Parental Explanations Support Children's Scientific Understanding

An explanation can be defined as talk that requests or makes a logical connection between objects, events, concepts, or conclusions (Beals, 1993). Explanations that meet this definition are relatively rare in everyday adult–child conversation (Rowe, 2012). Even when children ask questions that reference causal phenomena, parents respond with causal explanations only 50% of the time (Callanan and Oakes, 1992). In informal science settings, Kurkul et al. (unpublished) found that parents rarely produced explanations about electrical circuitry to their 4-year-old children unless prompted by a researcher. Additionally, Tabors et al. (2001) found that while interacting with their 5-year-old children around a magnet task, parents' *science process talk*, a broad category that contained explanations, comprised only 14% of their utterances compared to 55% of utterances that referred to superficial qualities of the magnets.

Furthermore, when parents do provide explanations, there is variation in the extent to which the explanation contains *features* that support children's learning (Baum et al., 2008; Russ et al., 2008; Corriveau and Kurkul, 2014; Mercier et al., 2014; Kurkul and Corriveau, 2018). The present study focuses on enhancing one feature of explanations: the presence of *mechanistic reasoning*. Mechanistic explanations, a form of causal explanations, provide information about the process through which a cause brings about an effect (Callanan and Oakes, 1992; Russ et al., 2008). Using a diary methodology, Callanan and Oakes (1992) found that parents most often used this type of explanation in response to 3- to 5-year-old children's inquiries about causal events. Mechanistic explanations are argued to be central to children's scientific theory-building because they highlight information that is not directly observable (Russ et al., 2008). Take, for instance, an explanation about how a light turns on. Although children may be able to observe a light turn on after a switch is flipped, they are not able to observe the process through which this occurs (i.e., electricity flows through the circuit to turn the light on). A more complete understanding can take place when an adult provides an explanation that elucidates the circuit mechanism.

Evidence that mechanistic explanations support children's science learning comes from several previous studies. For example, Nolan-Reyes et al. (2016) found that explanations that contained mechanistic reasoning about impossible and improbable events predicted 4- and 6-year-old children's possibility judgments and causal justifications for those judgments. Frazier et al. (2009) found that when experimenters provided 3- to 5-year-old children with mechanistic explanations, they were more likely to engage in subsequent information-seeking behaviors (i.e., asking follow-up questions) than when the experimenter provided a non-explanation. Studies with parents as the interlocutors also support the importance of mechanistic reasoning. For example, Willard et al. (2019) conducted a museum-based intervention study that encouraged parents to ask questions about gears in order to prompt mechanistic language (e.g., torque, motion) from their 4- to 6-year-old children (e.g., *How do the gears work?; What will happen when the gear moves?*). Parents who were assigned to the explanation condition had children who engaged in more scientific discourse and more often tested how the gears worked compared to a condition in which parents only encouraged children's exploration. Kurkul et al. (unpublished) extended Willard et al. (2019) findings by showing that mechanistic conversations about circuitry between parents and 4-year-old children predicted children's subsequent recall of the mechanistic explanation and their ability to transfer their knowledge to a novel scientific task.

Nevertheless, although children learn a great deal from explanations containing mechanistic reasoning, adults often struggle to adjust their language to provide accurate, yet developmentally appropriate mechanistic explanations, even if they understand the scientific mechanism under discussion (Gleason and Schauble, 1999; Schauble et al., 2002; Shtulman and Checa, 2012; Vlach and Noll, 2016). In the current study, we explored whether modeling mechanistic explanations using storybooks might increase parental use of these explanations to children during a subsequent informal science interaction.

To our knowledge, no study has examined whether parental explanations can be impacted via storybooks. However, other interventions without storybooks have proven effective and provide rationale for the current method. One strategy has been to provide parents with written instructions—termed "conversation cards"—for how to interact with their children (Fender and Crowley, 2007; Benjamin et al., 2010; Gutwill and Allen, 2010; Jant et al., 2014; Willard et al., 2019). In these studies, adult-child dyads or small groups of adults and children—mostly preschool and early elementary aged—receive different instructions for how to interact around a scientific exhibit. For example, Gutwill and Allen (2010) examined families' interactions in a museum exhibit by comparing control families (who received no intervention) with families who received a scientific intervention known as "Juicy Questions," which was aimed at increasing families' use of questions, investigations, and discussions. Families who received the Juicy Questions intervention asked more questions and engaged in more inquiry-based exploration than families in the control condition (Gutwill and Allen, 2010). Additionally, Jant et al. (2014) presented dyads comprising parents and their 4-year-old children with conversational cards prior to visiting an exhibit, finding that dyads who received cards with elaborative questions (i.e., those that encouraged a multiword response from children) engaged in conversation containing more elaborative talk as compared with those dyads who did not receive conversational cards. Finally, Willard et al. (2019) found that parent-child dyads assigned to an explanation condition in which parents were encouraged to ask their child questions about how a set of gears works, engaged in more mechanistic conversation about gear functions. Moreover, such talk predicted children's scientific exploration, operationalized as the testing of gears. Taken together, this research indicates it is possible to explicitly instruct adults to interact with children in certain ways (e.g., prompting parents to ask questions versus provide direct instruction or encouragement) in order to impact the scientific content of parent-child conversation and children's independent learning.

A second, similar method conducted primarily in museum settings is to invite experts (e.g., experimenters, teachers, or museum educators) to model explanations for parents. For example, in Marcus et al. (2017), researchers demonstrated how to build skyscrapers to parents and 5- to 6-year-old children and then provided explanations about building engineering (e.g., what supports help make buildings strong). Dyads who received engineering explanations were able to transfer knowledge to a subsequent building activity compared to dyads who only received a building demonstration. Such scaffolded interventions work well in certain settings—such as on the museum floor or in the classroom where experts can join into dyadic interactions to scaffold learning. However, a constraint of this method is that parent-child interactions are dependent on initial modeling by an experimenter or teacher who is present. Instead, the present study sought to determine whether storybooks could replace experts in modeling the use of mechanistic explanations for parent-child interactions.

Science Learning From Shared Book-Reading Interactions

Although typically examined in relation to early literacy outcomes (Fletcher and Reese, 2005) such as vocabulary (Wasik et al., 2016; Flack et al., 2018) and narrative ability (Zevenbergen and Whitehurst, 2003), book-reading interactions have also been used to examine the transmission of conceptual and scientific information between parents and preschool-aged children (Ganea et al., 2011), although note that this prior research did not examine transmission of mechanistic explanations. For example, manipulating book text to describe social categories using generic or non-generic language is associated with changes in parents' subsequent essentialist language (Gelman et al., 2004; Rhodes et al., 2012; Chalik and Rhodes, 2015). Further, varying the linguistic structures of the story text (e.g., syntactically complex phrases, future tense) impacts 4-year-old children's subsequent discourse and thinking (Vasilyeva et al., 2006; Leech et al., 2019a,b). Picture books hold a number of benefits for delivering scientific information to parent-child dyads. For example, the text is standardized to ensure accuracy of the scientific information, the content often contains information the dyad may not encounter frequently in the real world, and the setting provides an enjoyable situation where the parent and child share attention around an object (Ninio and Bruner, 1978; Kelemen et al., 2014). Previous work has described these scientific picture books as belonging to the *informational* or *expository* genre, which elicit more parent-child discourse, cognitively challenging questions, and opportunities to engage in reasoning than other genres (Duke, 2000; Anderson et al., 2004).

Current Study

Exposing children to mechanistic explanations through book reading may be an ideal context for promoting scientific discourse because of the opportunity to standardize aspects of the text and as a result, transmit accurate scientific information to the dyad, just as experts do on the museum floor or in the classroom. This method may be especially useful for transmitting scientific information to parents who may be unsure of how to adapt this information for their preschool-aged children.

For this study, we focused on explanations about electricity and the mechanism that makes electricity work. We invited parent-child dyads to read one of two storybooks about electricity, one that contained mechanistic explanations (e.g., "*It is a kind of energy that makes things move, light up or get hot*") or another that contained explanations that did not include mechanistic information (e.g., "*It's a kind of energy*

that we can't always see but it is very powerful"). These non-mechanistic explanations were primarily analogical or those that connected the concept of electricity to prior knowledge, which have also been found to facilitate preschool children's scientific understanding (Crowley and Jacobs, 2002; Valle and Callanan, 2006). Thus, dyads in both conditions read books that contained explanations that previous research has found to facilitate learning. After reading the book together, parents and children worked together during the *dyadic phase* to assemble electrical components of a circuit toy so that a light would turn on. Finally, during the *test phase*, children first independently assembled a novel circuit, and second, answered comprehension questions to gauge their understanding of electricity's mechanism.

Our first research question asked whether reading storybooks containing mechanistic language leads to an increase in (a) parent-child use of mechanistic language and (b) children's understanding of electricity's mechanism measured during the *test phase*. We predicted that dyads in the mechanistic condition would use more mechanistic language than those in the non-mechanistic condition. We also predicted that children in the mechanistic would outperform children in the non-mechanistic condition on outcome measures in the *test phase*. Our second research question examined whether the frequency of mechanistic language produced by dyads would *strengthen* (i.e., statistically moderate) the effect of reading mechanistic storybooks on children's scientific understanding. In line with social interactionist theories, conversations with more knowledgeable others help to organize children's knowledge and allow the child to access it in subsequent situations (Vygotsky, 1978). On this theory, exposure to mechanistic explanations in storybooks alone may not be sufficient to teach children about the mechanisms underlying scientific concepts. Rather, we predicted that additional conversation between parents and children may help reinforce the mechanistic explanations modeled in the storybook, allowing the child to retrieve this knowledge during the *test phase*.

MATERIALS AND METHODS

Participants

Seventy-six dyads were recruited to participate in the present study. Dyads were eligible for the study if they agreed to speak English during the study procedures and the child was between the ages of 4 and 6. Of the 76 dyads, we excluded $n = 7$ who were outside the age range and $n = 8$ who did not complete all procedures. This left a sample of 60 dyads for analysis. Note that $n = 2$ videos were not clear enough to transcribe and code conversation during the dyadic phase, leaving a sample of $n = 58$ for analyses that involve mechanistic conversation.

Our sample size was justified by an *a priori* power analysis using the G*power program (Faul et al., 2009). We powered the study to answer our second research question (whether dyadic conversation moderates the effect of storybook condition on children's scientific understanding) because the planned analyses required the most potential model parameters. The power analysis indicated that with 55 children, we have 0.80 power to detect a small effect size ($f^2 = 0.15$) for the moderation effect using a multiple regression model.

Participating children (21 girls, 39 boys) were on average 60.3 months ($SD = 5.99$; range = 49.3–71.5 months). Three parents declined to provide a date of birth but identified the child as being within the age group required for the study. The majority of parents were highly educated, with 95% receiving at least a 4-year college degree (note, $n = 6$ parents declined to provide their educational background). Thirty percent of the parent sample ($n = 17$) self-identified as working in a STEM field, as measured by the question, "Do you consider your primary occupation to be in a STEM (science, technology, engineering, math) field? Yes or No."

Procedure

Dyads were recruited on the floor of a science museum, a laboratory, or preschools all located in a large northeastern city in the United States. All families were drawn from a similar population. Participant dyads were randomly assigned to one of two book-reading conditions: mechanistic ($n = 32$) or non-mechanistic ($n = 28$). Child age and sex were balanced across conditions. As compensation, participants at the museum were given a sticker and participants at schools and the laboratory were given a book. Data were collected between June 2018 and December 2018.

The study consisted of four phases in a fixed order: a *pretest* to gauge children's preexisting understanding of electricity, a *book-reading interaction* during which dyads read their condition-specific storybook (mechanistic or non-mechanistic), a *dyadic phase* in which dyads completed a circuit task, and an individual child *test phase*. All phases were videotaped for later transcription, coding, and scoring. The entire procedure lasted approximately 15 min.

Pretest

Children were asked four questions about electricity in order to establish a measure of prior knowledge and to ensure the task was developmentally appropriate. First, children were asked: "*Have you ever heard of something called electricity?*" If the child answered yes, then the experimenter asked three follow-up questions: "*What do you think electricity is? Where does electricity come from? How does electricity travel?*" For the last two questions, the child was presented with a set of options, provided in pictorial form: "*Where does electricity come from?*" was asked with pictures of batteries, a faucet, and a planet (presented as outer space). "*How does electricity travel?*" was asked with pictures of wires, pipes, and a truck. The first two questions were asked in a fixed order; the last two questions were randomized.

Book-Reading Interaction

Parent-child dyads read a picture book containing explanations consistent with their condition. The experimenter provided the dyad with the condition-specific storybook and asked the parent to share the book with his or her child just like he or she would do at home. No time limit was given for completing the book and the average time spent on this phase was approximately 5 min.

A commercially available picture book, *Oscar and the Bird: A Book about Electricity* (Jeff Waring) was modified for the current study. Two versions of the text were prepared, which were matched on number of words, length of paragraphs, pictures, and linguistic complexity. Both books had 18 pages, 13 of which contained text and illustrations. The remaining five pages contained illustrations only. Books averaged 3.4 sentences per page. The critical difference between conditions was the content of the explanations. Each book contained eight explanations embedded into the text. For example, "*Electricity is a kind of energy that makes things move, light up, or get hot.*" The dyads in the non-mechanistic condition read a version in which the concepts were introduced in a procedural or analogical manner. For example, "*Electricity is a kind of energy that we can't always see but is very powerful.*"

Dyadic Phase

Immediately following the book-reading interaction, the dyad was presented with a SnapCircuit© board and corresponding pieces (**Figure 1**). The procedure was identical across conditions. The child was first asked whether he or she had seen a toy like this before in order to take into account prior familiarity with the stimulus. Thirteen dyads (21.7% of sample) reported familiarity with the toy. The experimenter then said, "*You can put the pieces together and you can take them apart. Do you think you could the pieces together to make the light turn on?*" Dyads were given approximately 5 min to complete the task, after which the experimenter intervened and introduced the next phase. Stimuli consisted of seven components: blue snap wire pieces, a battery holder containing three 1.5 V AA batteries, a 2.5 V light, and a press switch.

Book Reading and Dyadic Phase Measures

Circuit completion was measured during the dyadic phase by whether the dyad successfully constructed the circuit to turn the light on. We also calculated the number of minutes the dyad spent assembling the circuit.

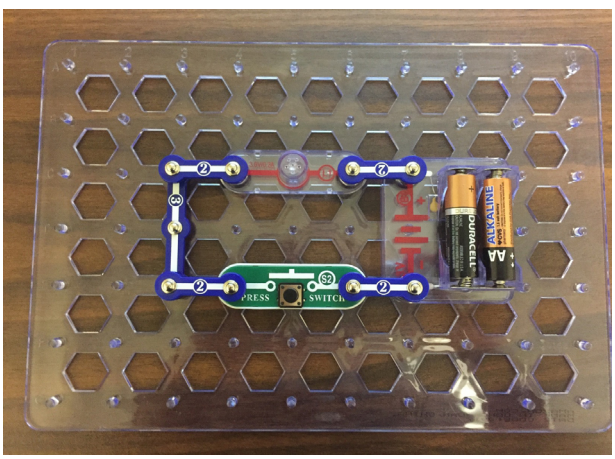


FIGURE 1 | Example of completed circuit used in dyadic interaction.

Second, we analyzed the conversations during the book reading and dyadic circuit phases to produce a measure of mechanistic discourse. First, all videos were transcribed at the level of the utterance by a team of research assistants according to the conventions of Child Language Data Exchange System (CHILDES; MacWhinney, 2000). Each transcript was then verified for accuracy by another trained research assistant. We then coded these transcripts for utterances that referenced mechanistic reasoning by adapting a coding scheme from Russ et al. (2008). Only extratextual speech (i.e., utterances that were not verbatim text) during the book-reading interaction was included in this measure. All mechanistic utterances during the dyadic phase were counted. Note that because mechanistic language was relatively rare, collapsing the two phases gave us more variation for the planned regression analyses. Reliability of two coders based on 15% of the transcripts resulted in 82% agreement (kappa = 0.80).

Our coding scheme captured different types of mechanistic language described by Russ et al. (2008): describing the target phenomenon (i.e., "*Make the light turn on*"); identifying entities by mentioning a circuit or connections (i.e., "*We have to connect the pieces in a circuit*"); identifying the activities and functions of the different entities (i.e., "*Electricity travels through the circuit to make the light turn on*"); and chaining backward (i.e., "*Why did the light turn on?*") or forwards (i.e., "*If we took a piece off, would the light still turn on?*") (**Table 1**). All utterances that fell into these categories were coded and summed to create a measure of mechanistic language. We chose to combine parental

TABLE 1 | Examples of mechanistic language during parent–child interaction.

Category	Definition	Examples
Describing target phenomenon	State phenomenon dyad is trying to produce	• We have to make the light turn on.
Identifying entities	Description of the enabling conditions that will make the mechanism run (i.e., light turn on)	• We have to connect the pieces in a circuit.
Identifying activities	The relevant activities that the entities engage in (functions of entities that cause changes in surrounding entities)	• Electricity travels through wires or the circuit. • Battery powers the circuit. • Switch controls the flow of electricity. • Wires connect the circuit.
Chaining backward/forward	A reasoning strategy that uses knowledge about the causal structure of the world to make claims about what must have happened previously (backward) or what will happen next (forward)	• Backwards: "Why did the light turn on?" • Forwards: "If we took a piece off, would the light still turn on?"

and child language together into one measure because we conceptualized the conversation as a co-construction of knowledge. For reference, **Table 2** displays mechanistic talk for both speaking partners separately. This table shows that the proportion of mechanistic utterances is equivalent for parents and children, indicating that both speaking partners contributed to the conversation.

Further, because time spent reading and interacting with the circuit varied across dyads, we created proportion measures by dividing dyads' mechanistic utterances by the total number of utterances produced during the interaction.

Test Phase

This phase consisted of two tasks: an independent circuit task and comprehension questions. While the child completed this phase, the parent was given a short paper-and-pencil survey, which consisted of demographic items, reading habits at home, and feedback on the current study.

Test Phase Measures

To determine if the child could independently complete the circuit activity, the researcher presented the child with a novel circuit and seven circuit components: differently sized wire pieces, batteries, a sliding switch, and a fan. The child was given approximately 5 min to put the pieces together to make the fan turn on. If the child did not complete the task in 5 min, the researcher suggested moving on, and offered to help finish putting the circuit together at the end of the study. Children's task completion was measured by their successful construction of a circuit that could turn the fan on within the time limit.

To measure children's comprehension of electrical circuitry, the child was shown two pictures of a circuit similar to those used in the previous phases (**Figures 2A,B**). **Figure 2A** displays an unsuccessful mechanism (i.e., the circuit is disconnected) and **Figure 2B** displays a successful mechanism (i.e., a connected circuit). The child was asked: "*If I were to press the button in this circuit, do you think the fan would turn on? Yes or no? Why?*" Children were presented with one picture at a time and the order of presentation was randomized across participants. We coded the answers children provided to the posttest comprehension task by counting the number of correct answers (out of two) children earned (0 = incorrect; 1 = correct).

TABLE 2 | Means, standard deviations, and ranges for mechanistic language variables separated by parent and child.

	Parent	Child
Mechanistic utterances	*M* = 9.90 (*SD* = 8.95; range = 0–35)	*M* = 1.26 (*SD* = 1.88; range = 0–9)
Total utterances	*M* = 160.31 (*SD* = 41.6; range = 83–258)	*M* = 25.69 (*SD* = 20.41; range = 1–117)
Proportion mechanistic utterances	*M* = 0.06 (*SD* = 0.05; range = 0–0.17)	*M* = 0.06 (*SD* = 0.10; range = 0–0.39)

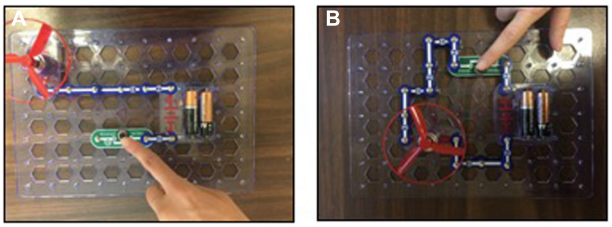


FIGURE 2 | Children's comprehension of electricity's mechanism was probed by asking whether a disconnected circuit **(A)** and connected circuit **(B)** would make the fan go.

RESULTS

First, we ensured that there were no differences in demographic characteristics between the two conditions. Neither parent education (p = 0.67) nor STEM occupational status (p = 0.59) differed significantly between the mechanistic and non-mechanistic conditions. Similarly, family reading habits—the frequency with which parents reported reading informational books such as *Oscar and the Bird*—were similar across the two conditions, χ^2 (n = 60, df = 4) = 7.11, p = 0.13. For reference, five parents reported reading informational books *hardly ever*, n = 9 reported *once or twice a month*, n = 37 reported *once or twice a week*, and n = 16 reported *almost daily*. We also examined children's preexisting knowledge of electricity using pretest data, finding that 75% of children in the non-mechanistic and 84% of children in the mechanistic condition reported hearing about electricity, χ^2 (n = 59, df = 1) = 0.72, p = 0.40. Finally, there was no significant condition difference in prior familiarity with the circuit stimuli, χ^2 (n = 59, df = 1) = 0.54, p = 0.46.

Do Mechanistic Storybooks Impact Dyadic Scientific Discourse?

Our first research question asked whether storybook condition led to differences in dyadic circuit assembly and mechanistic discourse. Looking first at the entire sample, n = 48 (80%) dyads successfully completed the joint circuit task. Importantly, condition differences emerged: nearly all dyads in the mechanistic condition (94%) constructed a circuit that turned on the light compared to 64% of dyads in the non-mechanistic condition, Fisher's exact test (n = 60, df = 1), p = 0.008. Of those who finished the task, the average time to complete the circuit was 3 min, 40 s (*SD* = 1:40.56; range = 0:53–7:19). Mechanistic condition dyads (mean time = 3:18) were no faster at completing the task than non-mechanistic dyads (mean time = 3:59), t(44.18) = 1.51, p = 0.14 (corrected for unequal variance across groups).

We then explored condition differences in dyads' use of mechanistic conversation. Collapsing both conditions (n = 58), dyads produced an average of 186.00 (*SD* = 52.40; range = 96–316) total utterances, of which 11.16 (*SD* = 9.65; range = 0–37) were coded as mechanistic. This corresponded to approximately 6% of dyads' total talk. A significant difference in mechanistic language by condition emerged: the proportion of mechanistic utterances was significantly greater for dyads who read

mechanistic storybooks (M = 0.07; SD = 0.05) than those who read non-mechanistic storybooks (M = 0.04; SD = 0.04), $t(56) = 2.41$, $p = 0.01$, Hedges' $g = 0.64$. Note there was no significant difference in the *total* number of utterances produced within the mechanistic (M = 182.56; SD = 50.82) and non-mechanistic conditions (M = 190.23; SD = 54.99), $t(56) = -0.55$, $p = 0.58$, Hedges' $g = 0.15$, suggesting that condition differences were isolated to mechanistic conversation only. Although dyads who completed the circuit (M = 0.06; SD = 0.05) used more mechanistic language than those who did not (M = 0.04; SD = 0.04), this difference was not statistically significant, $p = t(56) = 1.41$, $p = 0.17$. This lack of a statistical difference might be associated with the high percentage of dyads across both conditions (80%) who completed the joint circuit activity.

Is There a Relation Between the Storybook Manipulation and Children's Scientific Understanding?

We then examined the effects of condition on children's scientific understanding during the *test phase*. When asked to independently construct a circuit, 55% of children (n = 33) succeeded in turning on the fan. We found that 69% of children in the mechanistic condition constructed a circuit that turned on the fan as compared to 39% of the children in the non-mechanistic condition, χ^2 (n = 60, df = 1) = 5.24, $p = 0.02$. Children in the mechanistic condition (mean time = 3:20 min:s) were no faster at completing the circuit than children in the non-mechanistic condition (mean time = 3:33 min:s), $t(55) = 0.49$, $p = 0.63$. Further, children who succeeded at turning the fan on with their parents were more likely to have completed the dyadic circuit activity, χ^2 (n = 60, df = 1) = 8.91, $p = 0.003$.

Next, we turned to children's responses to the posttest comprehension items as a second index of their science understanding. Recall that comprehension was assessed by presenting children with pictures of a disconnected and connected circuit and asking them to reason about the outcome, that is, whether or not the fan would turn on. Forty-seven percent of children correctly answered that the disconnected circuit would not turn on, and 75% of children correctly answered that the connected circuit would turn on. An ordinal logistic regression model (Dependent Variable [DV] being 0 = neither picture correct, 1 = 1 picture correct, 2 = both pictures correct) indicated that children who read mechanistic storybooks were no more likely to correctly answer comprehension questions about the mechanism of electricity than children who read non-mechanistic storybooks, Estimate = 0.17 (SE = 0.49), Wald = 0.12, $p = 0.73$.

Does Dyadic Mechanistic Conversation Strengthen the Relation Between Mechanistic Storybooks and Children's Scientific Understanding?

Although there was no direct effect of Condition on children's posttest comprehension, we hypothesized this effect may be statistically moderated by children's social interactions with their parents. This hypothesis comes from our theory that language occurring during social interactions helps build children's knowledge and transfer it to novel situations.

We tested this hypothesis through a moderation analysis, regressing posttest comprehension scores on Condition, proportion of mechanistic utterances produced by the dyad (Mechanistic Conversation), and the interaction between the latter two terms (**Table 3**, model 1). We also controlled for whether the dyad successfully assembled the circuit (Dyadic Outcome) in order to isolate the effect of conversation. The overall model was significant, $R^2 = 0.19$, $F(4,52) = 3.07$, $p = 0.02$. Our main interest was in the Condition × Mechanistic Conversation interaction, which was also significant, $\Delta R^2 = 0.10$, $\Delta F(1,52) = 6.13$, $p = 0.01$, $b = 10.85$, $t(52) = 2.48$, $p = 0.01$. We probed the nature of the interaction by testing the conditional effect of Condition at three levels of mechanistic conversation: one standard deviation below the mean (proportion of utterances coded as mechanistic = 0.01), at the mean (0.06), and one standard deviation above the mean (0.10) (**Figure 3**). The Johnson–Neyman technique showed the relation between reading mechanistic storybooks and children's success on the comprehension task was significant when dyads' mechanistic talk comprised at least 11% of their total utterances during the circuit task. Thus, children in the mechanistic storybook condition were more likely to correctly answer the posttest comprehension questions when they participated in higher levels of mechanistic conversation with their parents. There was no

TABLE 3 | Regression models showing moderated effects of condition on children's science understanding.

Variable	Posttest comprehension model 1		Independent circuit task model 2	
	Coefficient	*t*	Coefficient	*z*
Intercept	1.35 [0.84, 1.87]	5.32***	−2.57 [−4.67, −0.47]	−2.40**
Dyadic outcome	−0.17 [−0.68, 0.34]	−0.67	2.03 [0.25, 3.80]	2.24*
Condition	−0.56 [−1.19, 0.07]	−1.79 ∼	1.60 [−0.38, 3.58]	−1.58
Mechanistic conversation	−2.26 [−9.37, 4.85]	−0.64	16.56 [−9.93, 43.05]	1.22
Condition × Mechanistic conversation	10.85 [2.06, 19.63]	2.48**	−18.34 [−49.52, 12.84]	−1.15
Model fit	R^2 19.1%		−2LL 66.80	

*∼p < 0.10; *p < 0.05; **p < 0.01; ***p < 0.001; dyadic outcome = whether dyad successfully completed the dyadic circuit task; brackets display 95% confidence intervals for model coefficients.*

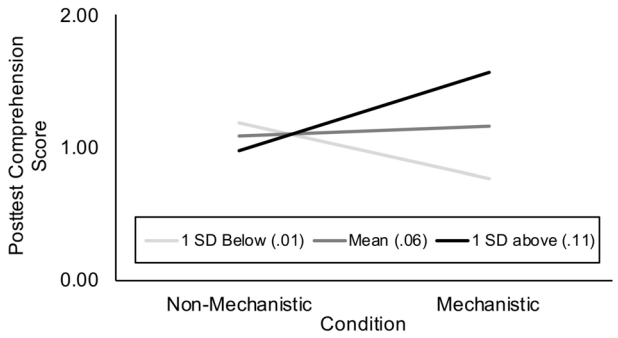


FIGURE 3 | Mechanistic language moderates the effect of condition on children's comprehension. Lines refer to the proportion of mechanistic utterances produced by dyads. Children in the mechanistic storybook condition were more likely to correctly answer the posttest comprehension questions only when they participated in higher levels of mechanistic conversation with their parents (black line).

such magnifying effect of mechanistic conversation for children who read storybooks with non-mechanistic explanations.

Our final analysis examined whether mechanistic conversation also moderated the condition effect on children's independent circuit task performance. **Table 3** (model 2) shows the results of a logistic regression predicting circuit success from Condition, Mechanistic Conversation, and the interaction between the latter two terms. This model fits better than a constant only model, χ^2 (df = 4) = 12.98, p = 0.01. However, we did not find evidence that conversation served a moderating role, as indicated by a non-significant interaction effect, $b = -18.34$, z = 1.15, p = 0.25.

DISCUSSION

In this study, we examined how science storybooks may augment parent–child science conversations and children's learning of a scientific mechanism. In particular, our first research question explored whether embedding explanations about causal mechanisms into a storybook about electricity could affect (a) parent–child talk about electrical mechanisms and (b) children's individual understanding of such concepts. Drawing from social-interactionist theories, our second research question sought to determine whether mechanistic conversation produced spontaneously by parents and children magnified the effect of the storybook manipulation on children's independent scientific understanding during the *test phase*.

Does Embedding Explanations About Causal Mechanisms Into Storybooks Affect Parent–Child Science Understanding and Discourse?

Regarding (a) in the first research question, we found evidence that the book-reading manipulation was effective: dyads in the mechanistic condition were significantly more successful at completing a circuit task compared to dyads who read stories containing non-mechanistic explanations. Further, there was a significant difference in mechanistic language by condition, with dyads in the mechanistic condition using a greater proportion of mechanistic utterances compared to those in the non-mechanistic condition. Thus, it appears that providing verbal models of mechanistic reasoning via storybooks transfers to dyads' joint understanding of scientific mechanisms as well as talk about such concepts.

To our knowledge, this is the first study to use storybooks to model mechanistic explanations about science concepts. We argue that embedding mechanistic language into the storybook text provided parents with a developmentally appropriate model to use with their children. As many parents may understand the concept of electrical circuitry but struggle to explain it to their preschool-aged children (Gleason and Schauble, 1999; Schauble et al., 2002; Shtulman and Checa, 2012; Vlach and Noll, 2016), storybooks may be an effective method for scaffolding parents' explanations to children.

These results add to previous work showing how providing developmentally appropriate models of scientific explanations can help parents integrate this language into their conversations with children. For example, Jant et al. (2014) reported on a study where experimenters modeled explanations about building engineering, finding that modeling specific language for parents can modify the content of conversation and lead to deeper exploration and understanding of the topics at hand. Additionally, Benjamin et al. (2010) found that providing instructions prompting parents to ask more causal *wh*-questions during museum interactions was effective in boosting the use of such questions to children. The current study builds on this prior research by utilizing storybooks instead of museum educators or researchers as an interactive tool for prompting parental use of mechanistic language. Using a storybook method for modeling explanations may prove especially useful for scaling up parent–child interventions in the future, as a live experimenter or educator is not required to be present.

Moreover, most previous research on parent–child interventions around scientific discourse largely involves explicit instruction for parent–child interaction. For instance, previous studies have used conversational cards to explicitly instruct parents to ask more questions, be more elaborative, or encourage children's exploration (Benjamin et al., 2010; Jant et al., 2014; and see Boland et al., 2003 for an example in a non-scientific setting). In the current study, we employed a less explicit method of storybooks to model a type of explanation known to facilitate learning. We argue that the storybook delivery method can be seen as a strengths-based approach to fostering scientific discourse. Storybooks containing embedded explanations provide parents with accurate scientific information while also allowing for considerable latitude in how the parent chooses to use the storybook text based on individual family dynamics or cultural values (e.g., Gutiérrez and Rogoff, 2003; Kline, 2015). For instance, a parent may choose to read the text verbatim, may augment the text with subsequent discussion, or employ a combination of these practices. The adoption of such a strengths-based approach holds important implications for adapting this study

to other populations, as we discuss in more detail in the text that follows.

Interestingly, when examining the amount of mechanistic language produced in dyadic conversations, we found that it was quite rare, comprising approximately 6% of dyads' total utterances. Though this proportion may appear small, it is consistent with prior research. For instance, Kurkul et al. (unpublished) found that parents do not often produce spontaneous mechanistic explanations to their children, and observational studies that have measured the frequency of child-directed explanations report similar proportions (Beals, 2001; Rowe, 2012). Though explanatory language is relatively rare in everyday conversation (Rowe, 2012), research has shown that even infrequent participation in such conversations is positively related to children's scientific understanding (Tabors et al., 2001).

Does Embedding Explanations About Causal Mechanisms Into Storybooks Affect Children's Science Learning?

Regarding (b) in the first research question, we observed that storybook condition affected children's learning during the independent circuit task. Here, children in the mechanistic condition were more successful at completing the circuit task than children in the non-mechanistic condition. These results show that not only does a book-reading manipulation affect parent–child scientific discourse, it also impacts children's scientific behaviors on a different, but related task. This finding provides further evidence and extends Kurkul et al. (unpublished), where children's performance on a similar circuit task was enhanced when they were systematically exposed to mechanistic explanations from experimenters. Here we show that using storybook interactions with parents is similarly effective. Because prior research has often looked solely at storybooks without a hands-on application task or solely at targeted games such as gears or circuits (Legare and Lombrozo, 2014; Kurkul et al., unpublished), our study extends this research by targeting both of these elements together.

Although we found that storybook condition influenced children's performance on the independent circuit task, our results indicated that there was no main effect of condition on children's responses to a series of comprehension questions gauging their mechanistic understanding. One explanation for this null result is that the comprehension questions relied on a deeper understanding of electricity's mechanism than the independent circuit task. It is possible that in order to complete the independent circuit, children may have simply relied on their procedural memories, replicating what they did with their parent in the dyadic circuit task. However, in order for children to correctly answer the comprehension questions, they likely needed to possess an accurate understanding of the underlying mechanistic concept. On this hypothesis, children may have needed more scaffolding—such as participation in mechanistic conversation with their parents—to fully acquire an understanding of electricity's mechanism.

Indeed, in our second research question, we found support for this hypothesis by showing that the effect of storybook condition on children's comprehension was statistically moderated by children's social interactions with their parents. The results indicate that mechanistic conversation amplified the effect of mechanistic storybooks on children's science learning. We interpret these results using a social-interactionist framework: reading stories that contain mechanistic language may not be sufficient to teach children about the mechanisms underlying scientific concepts. Instead, we argue that social interaction serves as the process for learning and therefore, helped to strengthen the mechanistic concepts conveyed in the storybook. In other words, the more the dyad used mechanistic language, the better the child performed on the comprehension assessment.

This finding extends prior research and provides further evidence that social interaction is an important process by which adults influence children's learning (Gunderson and Levine, 2011; Pruden et al., 2011; Marcus et al., 2017). We reached a similar conclusion to Jant et al. (2014) in that children learn about scientific mechanisms through a combination of "doing and talking." Indeed, children can learn some of the information regarding electrical circuitry through hands-on contact with the circuit game. However, a more complete understanding of the mechanism underlying the circuit seems to require conversation and interaction with more knowledgeable others. This conclusion draws upon and unites two relatively separate literatures on how children learn, one of which focuses on children's hands-on exploration (e.g., Piaget, 1964) and another that emphasizes the role of conversation and interaction (e.g., Vygotsky, 1978). Taken together, these findings indicate that systematic exposure to mechanistic explanations via storybooks coupled with opportunities to discuss and explore the mechanisms under question can be an effective way to improve children's understanding of scientific concepts.

Limitations

It is important to note that because data were collected primarily in a museum and laboratory context, the sample was drawn from a relatively educated, higher socioeconomic population of parents. One important direction for future research is to determine whether this manipulation would be successful with other populations of parent–child dyads, such as families from lower socioeconomic backgrounds. We argue that our storybook delivery method—as opposed to an explicit intervention with an experimenter—may potentially increase parents' willingness to adopt the explanatory style modeled in the storybook. However, it is important to note that previous studies have found that higher socioeconomic status caregivers, particularly those with more education, report reading more frequently (Bus et al., 1995) and using more explanations with preschool-aged children (Rowe, 2012). These socioeconomic status differences in baseline patterns of parent–child interaction may suggest a need to modify the current procedures, by for example, providing more scaffolding during the book-reading interaction. Nevertheless, the application of a scientific storybook intervention across populations is an important question for future research.

Second, we chose to use parents rather than experimenters as children's conversational partner to create a more naturalistic

interaction. This design choice may have resulted in additional unexplained variation in features of conversation than if we used an experimenter. For instance, parents' use of other linguistic features such as questions or non-verbal scaffolds may have also contributed to children's science learning. However, because dyads were randomly assigned to the conditions and mechanistic explanations were the only manipulation, other conversational features that contribute to children's learning should be equally distributed across the two conditions. Though examining potential condition differences in these features was beyond the scope of the study, we found no difference in dyads' overall talkativeness (i.e., number of total utterances) across conditions. This gives us some evidence that it was the increase in mechanistic conversation, not other features of conversation, that caused improvements to children's individual and joint learning.

Further, because parents did not implement the intervention in a standardized way (as an experimenter would have), there was wide variation in the amount of mechanistic language children heard, even within the same condition. This variation is evidenced by large standard deviations in the mechanistic language variables. Variation could possibly reflect that the "uptake" of the intervention was more successful for some dyads more than others. Future research with a larger sample size could test this claim by examining who this intervention works best for and why.

Implications and Future Directions

There are several future directions that stem from this work. First, we focused on the concept of electricity and the mechanism that makes electricity work because it is developmentally appropriate for preschool-aged children and could be discussed during this single timepoint study. However, future studies may consider embedding other STEM concepts into storybooks, particularly those that are more opaque, such as germs, or more complex, such as forces and gravity. Another future direction is to determine the longitudinal effects of scientific storybooks on discourse and evaluate the longevity of the knowledge the children acquired during the intervention. As parents were asked to read the storybook with children on a single occasion, a next step could be to provide families with a set of books to take home and examine whether longer-term exposure leads to larger gains in children's scientific understanding. Indeed, we know that repeated reading of storybooks leads to more extratextual talk and engagement from the adult and child (Robbins and Ehri, 1994; Senechal et al., 1995; Leung, 2008).

Our study has the potential to inform research on parent–child interaction and science learning. First, the results from this study demonstrate that subtle differences in how a concept is presented, namely, the presence or absence of mechanistic reasoning, can influence children's learning. Recall that the illustrations, accuracy of information, linguistic complexity of text, and the length of the storybook explanations were equivalent across the two conditions. Thus, we are able to isolate changes in discourse and learning to the mechanistic manipulation. This finding holds implications for educators, parents, and other caregivers regarding the importance of integrating mechanistic reasoning into informal and formal educational settings. Second, storybooks may provide a window into science content that parents are less comfortable talking about with their child. Presenting scientific information in a storybook containing both narrative and informational elements may make the information more accessible, interesting and appealing to both parents and children, and is a cost effective and fun way to join learning with leisure. In considering formal learning contexts, this study can inform educators about ways to enhance science instruction across the curriculum. Research indicates that only 3–11% of preschoolers' time is spent on science activities in early childhood classrooms versus nearly double the amount spent on literacy activities (La Paro et al., 2009; Piasta et al., 2014). Our results illustrate that literacy and science can be integrated, where children learn science content knowledge by engaging shared book reading. Indeed, shared reading of informational books is a practice that many educators utilize frequently.

CONCLUSION

In sum, embedding mechanistic explanations in storybooks can be an effective way to increase children's science discourse and learning. We present a framework for enhancing parent–child interactions that can be implemented in informal and formal learning settings by a variety of caregivers. Our findings add to the existing evidence that conversation between an adult and child plays an essential role in the development of mechanistic reasoning and more generally their understanding of science concepts during the early childhood period.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Boston University IRB. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

KL, YJ, and KC designed the study. KL, AH, and YJ collected and processed the data. KL and AH analyzed the data and co-wrote the manuscript. KC and YJ provided feedback on manuscript drafts. All authors approved the final manuscript draft.

FUNDING

This research was supported by grants #1652224 and #1640816 from the National Science Foundation to KC.

REFERENCES

Anderson, J., Anderson, A., Lynch, J., and Shapiro, J. (2004). Examining the effects of gender and genre on interactions in shared book reading. *Liter. Res. Instruct.* 43, 1–20. doi: 10.1080/19388070409558414

Baum, L. A., Danovitch, J. H., and Keil, F. C. (2008). Children's sensitivity to circular explanations. *J. Exp. Child Psychol.* 100, 146–155. doi: 10.1016/j.jecp.2007.10.007

Beals, D. E. (1993). Explanatory talk in low-income families' mealtime conversations. *Appl. Psycholing.* 14, 489–513. doi: 10.1017/s0142716400010717

Beals, D. E. (2001). "Eating and reading: links between family conversations with preschoolers and later language and literacy," in *Beginning Literacy With Language*, eds D. K. Dickinson and P. O. Tabors (Baltimore, MD: Brookes), 75–92.

Beals, D. E., and Snow, C. E. (1994). Thunder is when the angels are upstairs bowling: narratives and explanations at the dinner table. *J. Narrat. Life Hist.* 4, 331–352. doi: 10.1075/jnlh.4.4.06thu

Benjamin, N., Haden, C. A., and Wilkerson, E. (2010). Enhancing building, conversation, and learning through caregiver-child interactions in a children's museum. *Dev. Psychol.* 46, 502–515. doi: 10.1037/a0017822

Boland, A. M., Haden, C. A., and Ornstein, P. A. (2003). Boosting children's memory by training mothers in the use of an elaborative conversational style as an event unfolds. *J. Cogn. Dev.* 4, 39–65. doi: 10.1080/15248372.2003.9669682

Bruner, J. (1983). *Child's Talk: Learning to Use Language*. New York: W.W. Norton.

Bus, A. G., Van Ijzendoorn, M. H., and Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: a meta-analysis on intergenerational transmission of literacy. *Rev. Educ. Res.* 65, 1–21. doi: 10.3102/00346543065001001

Callanan, M. A., and Jipson, J. L. (2001). "Explanatory conversations and young children's developing scientific literacy," in *Designing for Science: Implications From Everyday, Classroom, And Professional Settings*, eds K. Crowley, C. D. Schunn, and T. Okada (Lawrence: Erlbaum Associates Publishers), 21–49.

Callanan, M. A., and Oakes, L. M. (1992). Preschoolers questions and parents' explanations: causal thinking in everyday activity. *Cogn. Dev.* 7, 213–233. doi: 10.1016/0885-2014(92)90012-g

Chalik, L., and Rhodes, M. (2015). The communication of Naïve theories of the social world in parent–child conversation. *J. Cogn. Dev.* 16, 719–741. doi: 10.1080/15248372.2014.949722

Corriveau, K. H., and Kurkul, K. E. (2014). "Why does rain fall?": Children prefer to learn from an informant who uses noncircular explanations. *Child. Dev.* 85, 1827–1835. doi: 10.1111/cdev.12240

Crowley, K., Callanan, M., Jipson, J. L., Galco, J., Topping, K., and Shrager, J. (2001a). Shared scientific thinking in everyday parent—child activity. *Sci. Educ.* 85, 712–732. doi: 10.1002/sce.1035

Crowley, K., Callanan, M. A., Tenenbaum, H. R., and Allen, E. (2001b). Parents explain more often to boys than to girls during shared scientific thinking. *Psychol. Sci.* 12, 258–261. doi: 10.1111/1467-9280.00347

Crowley, K., and Jacobs, M. (2002). "Building islands of expertise," in *Learning Conversations In Museum*, eds G. Leinhardt, K. Crowley, and K. Knutson (Mahwah, NJ: Lawrence Erlbaum), 333–356.

Crowley, K., and Siegler, R. S. (1999). Explanation and generalization in young children's strategy learning. *Child. Dev.* 70, 304–316. doi: 10.1111/1467-8624.00023

Duke, N. K. (2000). 3.6 minutes per day: the scarcity of informational texts in first grade. *Read. Res. Q.* 35, 202–224. doi: 10.1598/RRQ.35.2.1

Faul, F., Erdfelder, E., Buchner, A., and Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/brm.41.4.1149

Fender, J. G., and Crowley, K. (2007). How parent explanation changes what children learn from everyday scientific thinking. *J. Appl. Dev. Psychol.* 28, 189–210. doi: 10.1016/j.appdev.2007.02.007

Flack, Z. M., Field, A. P., and Horst, J. S. (2018). The effects of shared storybook reading on word learning: a meta-analysis. *Dev. Psychol.* 54, 1334–1346. doi: 10.1037/dev0000512

Fletcher, K. L., and Reese, E. (2005). Picture book reading with young children: a conceptual framework. *Dev. Rev.* 25, 64–103. doi: 10.1016/j.dr.2004.08.009

Frazier, B. N., Gelman, S. A., and Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult–child conversation. *Child. Dev.* 80, 1592–1611. doi: 10.1080/15248372.2015.1098649

Ganea, P. A., Ma, L., and DeLoache, J. S. (2011). Young children's learning and transfer of biological information from picture books to real animals. *Child. Dev.* 82, 1421–1433. doi: 10.1111/j.1467-8624.2011.01612.x

Gelman, R., Brenneman, K., Macdonald, G., and Roman, M. (2010). *Preschool Pathways To Science: Ways Of Doing, Thinking, Communicating And Knowing About Science*. Baltimore, MD: Brookes Publishing.

Gelman, S. A., Taylor, M. G., and Nguyen, S. P. (2004). Mother-child conversations about gender: understanding the acquisition of essentialist beliefs. *Monogr. Soc. Res. Child Dev.* 69, 116–127. doi: 10.1111/j.1540-5834.2004.06901001.x

Gleason, M. E., and Schauble, L. (1999). Parents' assistance of their children's scientific reasoning. *Cogn. Instruct.* 17, 343–378. doi: 10.1207/s1532690xci1704_1

Gunderson, E. A., and Levine, S. C. (2011). Some types of parent number talk count more than others: relations between parents' input and children's cardinal-number knowledge. *Dev. Sci.* 14, 1021–1032. doi: 10.1111/j.1467-7687.2011.01050.x

Gutiérrez, K. D., and Rogoff, B. (2003). Cultural ways of learning: individual traits or repertoires of practice. *Educ. Res.* 32, 19–25. doi: 10.3102/0013189x032005019

Gutwill, J. P., and Allen, S. (2010). *Group Inquiry at Science Museum Exhibits: Getting Visitors to Ask Juicy Questions*. San Francisco: Exploratorium.

Haden, C. (2010). Talking about science in museums. *Child Dev. Perspect.* 4, 62–67. doi: 10.1111/j.1750-8606.2009.00119.x

Harris, P. L., and Corriveau, K. H. (2014). "Learning from testimony about religion and science," in *Trust and Skepticism*, eds E. J. Robinson and S. Einav (London: Psychology Press), 36–49.

Harris, P. L., Koenig, M. A., Corriveau, K. H., and Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annu. Rev. Psychol.* 69, 251–273. doi: 10.1146/annurev-psych-122216-011710

Jant, E. A., Haden, C. A., Uttal, D. H., and Babcock, E. (2014). Conversation and object manipulation influence children's learning in a museum. *Child. Dev.* 85, 2029–2045. doi: 10.1111/cdev.12252

Jipson, J. L., Gülgöz, S., and Gelman, S. A. (2016). Parent–child conversations regarding the ontological status of a robotic dog. *Cogn. Dev.* 39, 21–35. doi: 10.1016/j.cogdev.2016.03.001

Kelemen, D., Emmons, N. A., Schillaci, R. S., and Ganea, P. (2014). Young children can be taught basic natural selection using a picture-storybook intervention. *Psychol. Sci.* 25, 893–902. doi: 10.1177/0956797613516009

Kline, M. A. (2015). How to learn about teaching: an evolutionary framework for the study of teaching behavior in humans and other animals. *Behav. Brain Sci.* 38:e31. doi: 10.1017/S0140525X14000090

Kurkul, K. E., and Corriveau, K. H. (2018). Question, explanation, follow-up: a mechanism for learning from others? *Child Dev.* 89, 280–294. doi: 10.1111/cdev.12726

La Paro, K. M., Hamre, B. K., Locasale-Crouch, J., Pianta, R. C., Bryant, D., Early, D. M., et al. (2009). Quality in kindergarten classrooms: observational evidence for the need to increase children's learning opportunities in early education classrooms. *Early Educ. Dev.* 20, 657–692. doi: 10.1080/10409280802541965

Lagattuta, K. H., and Wellman, H. M. (2002). Differences in early parent-child conversations about negative versus positive emotions: implications for the development of psychological understanding. *Dev. Psychol.* 38, 564–580. doi: 10.1037//0012-1649.38.4.564

Leech, K. A., Haber, A. S., Arunachalam, S., Kurkul, K., and Corriveau, K. H. (2019a). On the malleability of selective trust. *J. Exp. Child Psychol.* 183, 65–74. doi: 10.1016/j.jecp.2019.01.013

Leech, K. A., Leimgruber, K., Warneken, F., and Rowe, M. L. (2019b). Conversation about the future self improves preschoolers' prospection abilities. *J. Exp. Child Psychol.* 181, 110–120. doi: 10.1016/j.jecp.2018.12.008

Legare, C. H. (2014). The contributions of explanation and exploration to Children's scientific reasoning. *Child Dev. Perspect.* 8, 101–106. doi: 10.1111/cdep.12070

Legare, C. H., and Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *J. Exp. Child Psychol.* 126, 198–212. doi: 10.1016/j.jecp.2014.03.001

Leung, C. B. (2008). Preschoolers' acquisition of scientific vocabulary through repeated read-aloud events, retellings, and hands-on science activities. *Read. Psychol.* 29, 165–193. doi: 10.1080/02702710801964090

Lombrozo, T., Bonawitz, E. B., and Scalise, N. R. (2018). Young children's learning and generalization of teleological and mechanistic explanations. *J. Cogn. Dev.* 19, 220–232. doi: 10.1080/15248372.2018.1427099

MacWhinney, B. (2000). *The Childes Project: The Database*, Vol. 2. London: Psychology Press.

Marcus, M., Haden, C. A., and Uttal, D. H. (2017). STEM learning and transfer in a children's museum and beyond. *Merrill Palmer Q.* 63, 155–180.

Mercier, H., Bernard, S., and Clement, F. (2014). Early sensitivity to arguments: how preschoolers weight circular arguments. *J. Exp. Child Psychol.* 125, 102–109. doi: 10.1016/j.jecp.2013.11.011

Ninio, A., and Bruner, J. (1978). The achievement and antecedents of labelling. *J. Child Lang.* 5, 1–15. doi: 10.1017/s0305000900001896

Nolan-Reyes, C., Callanan, M. A., and Haigh, K. A. (2016). Practicing possibilities: parents' explanations of unusual events and children's possibility thinking. *J. Cogn. Dev.* 17, 378–395. doi: 10.1080/15248372.2014.963224

Piaget, J. (1964). Part I: cognitive development in children: piaget development and learning. *J. Res. Sci. Teach.* 2, 176–186. doi: 10.1002/tea.3660020306

Piasta, S. B., Pelatti, C. Y., and Miller, H. L. (2014). Mathematics and science learning opportunities in preschool classrooms. *Early Educ. Dev.* 25, 445–468. doi: 10.1080/10409289.2013.817753

Pruden, S. M., Levine, S. C., and Huttenlocher, J. (2011). Children's spatial thinking: does talk about the spatial world matter? *Dev. Sci.* 14, 1417–1430. doi: 10.1111/j.1467-7687.2011.01088.x

Ramani, G. B., Rowe, M. L., Eason, S. H., and Leech, K. A. (2015). Math talk during informal learning activities in Head Start families. *Cogn. Dev.* 35, 15–33. doi: 10.1016/j.cogdev.2014.11.002

Rhodes, M., Leslie, S. J., and Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proc. Natl. Acad. Sci. U.S.A.* 109, 13526–13531. doi: 10.1073/pnas.1208951109

Robbins, C., and Ehri, L. C. (1994). Reading storybooks to kindergartners helps them learn new vocabulary words. *J. Educ. Psychol.* 86, 54–64. doi: 10.1037/0022-0663.86.1.54

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development: child-directed speech and vocabulary. *Child Dev.* 83, 1762–1774. doi: 10.1111/j.1467-8624.2012.01805.x

Russ, R. S., Scherr, R. E., Hammer, D., and Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: a framework for discourse analysis developed from philosophy of science. *Sci. Educ.* 92, 499–525. doi: 10.1002/sce.20264

Schauble, L., Gleason, M., Lehrer, R., Bartlett, K., Petrosino, A., Allen, A., et al. (2002). "Supporting science learning in museums," in *Learning Conversations In Museums*, eds G. Leinhardt, K. Crowley, and K. Knutson (Mahwah, NJ: Lawrence Erlbaum Associates Inc), 425–452.

Senechal, M., Thomas, E., and Monker, J. A. (1995). Individual differences in 4-yearold children's acquisition of vocabulary during storybook reading. *J. Educ. Psychol.* 87, 218–229. doi: 10.1037/0022-0663.87.2.218

Shtulman, A., and Checa, I. (2012). Parent-child conversations about evolution in the context of an interactive museum display. *Intern. Electron. J. Element. Educ.* 5, 27–46.

Snow, C., and Kurland, B. (1996). "Sticking to the point: talk about magnets as a context for engaging in scientific discourse," in Discourse, Learning, and Schooling, ed D. Hicks (Cambridge: Cambridge University Press), 189–220. doi: 10.1017/CBO9780511720390.007

Tabors, P. O., Roach, K. A., and Snow, C. E. (2001). "Home language and literacy environment: final results," in *Beginning Literacy With Language: Young Children Learning At Home And School*, eds D. K. Dickinson and P. O. Tabors (Baltimore, MD: Paul H Brookes Publishing), 111–138.

Tenenbaum, H. R., Snow, C. E., Roach, K. A., and Kurland, B. (2005). Talking and reading science: longitudinal data on sex differences in mother–child conversations in low-income families. *J. Appl. Dev. Psychol.* 26, 1–19. doi: 10.1016/j.appdev.2004.10.004

Valle, A., and Callanan, M. A. (2006). Similarity comparisons and relational analogies in parent-child conversations about science topics. *Merrill Palmer Q.* 52, 96–124. doi: 10.1353/mpq.2006.0009

Vasilyeva, M., Huttenlocher, J., and Waterfall, H. (2006). Effects of language intervention on syntactic skill levels in preschoolers. *Dev. Psychol.* 42, 164–174. doi: 10.1037/0012-1649.42.1.164

Vlach, H. A., and Noll, N. (2016). Talking to children about science is harder than we think: characteristics and metacognitive judgements of explanations provided to children and adults. *Metacogn. Learn.* 11, 317–338. doi: 10.1007/s11409-016-9153-y

Vygotsky, L. S. (1978). *Mind in Society: The Development Of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Wasik, B. A., Hindman, A. H., and Snell, E. K. (2016). Book reading and vocabulary development: a systematic review. *Early Childh. Res. Q.* 37, 39–57. doi: 10.1016/j.ecresq.2016.04.003

Willard, A. K., Busch, J. T. A., Cullum, K. A., Letourneau, S. M., Sobel, D. M., Callanan, M., et al. (2019). Explain this, explore that: a study of parent–child interaction in a children's museum. *Child Dev.* 90, 598–617. doi: 10.1111/cdev.13232

Zevenbergen, A. A., and Whitehurst, G. J. (2003). "Dialogic reading: a shared picture book reading intervention for preschoolers," in *Center for Improvement of Early Reading Achievement, CIERA. On Reading Books To Children: Parents And Teachers*, eds A. van Kleeck, S. A. Stahl, and E. B. Bauer (New Jersey: Lawrence Erlbaum Associates Publishers), 177–200.

Ziv, M., Smadja, M. L., and Aram, D. (2013). Mothers' mental-state discourse with preschoolers during storybook reading and wordless storybook telling. *Early Childh. Res. Q.* 28, 177–186.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Leech, Haber, Jalkh and Corriveau. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 12 June 2020
doi: 10.3389/fpsyg.2020.01140

Theory of Mind, Personal Epistemology, and Science Learning: Exploring Common Conceptual Components

Natassa Kyriakopoulou[1]* *and Stella Vosniadou*[2]

[1] Department of Early Childhood Education, National and Kapodistrian University of Athens, Athens, Greece, [2] College of Education, Psychology and Social Work, Flinders University, Adelaide, SA, Australia

We investigated the hypothesis that theory of mind (ToM) and epistemological understanding promote the aspect of science learning that concerns the ability to understand that there can be more than one representation of the same phenomenon in the physical world. Sixty-three students ranging in age from 10 to 12 years were administered two false-belief ToM tasks, an epistemological understanding task that investigated beliefs about the nature of science and a science learning task. The science learning task required distinguishing and reflecting upon phenomenal and scientific depictions of phenomena in observational astronomy. A three-stage hierarchical multiple regression showed that ToM was a significant predictor of performance in the astronomy task, supporting the hypothesis of a common underlying conceptual component. The results also showed that performance in the personal epistemology–nature of science task was a significant predictor of performance in the astronomy task, even when ToM and age were taken into consideration. The results indicate that both ToM and epistemological understanding promote the ability to construct and reflect on phenomenal and scientific representations of the same situation in the physical world and have important implications for science education.

Keywords: theory of mind, personal epistemology, science learning, conceptual change, observational astronomy

OPEN ACCESS

Edited by:
Kai S. Cortina,
University of Michigan, United States

Reviewed by:
Sergio Morra,
University of Genoa, Italy
Garvin Brod,
Leibniz Institute for Research and Information in Education (DIPF), Germany

***Correspondence:**
Natassa Kyriakopoulou
ankyriak@ecd.uoa.gr

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *31 January 2020*
Accepted: *04 May 2020*
Published: *12 June 2020*

Citation:
Kyriakopoulou N and Vosniadou S (2020) Theory of Mind, Personal Epistemology, and Science Learning: Exploring Common Conceptual Components. Front. Psychol. 11:1140. doi: 10.3389/fpsyg.2020.01140

INTRODUCTION

The learning of science is a complex task for many students, requiring the development of a host of interrelated thinking skills and conceptual changes (Carey, 1985; Chi, 2009; Kuhn, 2011; Vosniadou, 2019). The present research focuses on one aspect of this development, which has to do with the ability to construct and flexibly manipulate more than one representation of the same phenomenon in the world. For example, it has been shown that between the ages of 6 and 12, children recategorize their concept of the Earth from that of a physical object (a flat ground with the Sun and Moon in the sky above) to an astronomical object (a rotating sphere, a planet revolving around the Sun; Vosniadou and Brewer, 1992, 1994; Vosniadou et al., 2004; Vosniadou and Skopeliti, 2014). This recategorization suggests that the children had constructed a second, scientific representation of the Earth, which was distinct from their original, perception-based representation.

Recent research has shown that scientific representations do not supplant initial, phenomenal representations but coexist with them (Shtulman and Valcarcel, 2012; DeWolf and Vosniadou, 2015; Vosniadou et al., 2018). One of the important tasks of science learning and scientific thinking is to learn to navigate flexibly between such alternative representations (Pozo et al., 1999; Schwartz and Brown, 2013; Treagust et al., 2017). This is not an easy task. In a series of experiments, Kyriakopoulou and Vosniadou (Kyriakopoulou and Vosniadou, 2004, 2008; Vosniadou and Kyriakopoulou, 2006) presented elementary school children with phenomenal and scientific depictions of the same astronomical phenomena, such as the shape of the Earth, the structure of the solar system, and the day/night cycle, and asked them to select those that were closer to the way "things appear to be" and those that were closer to the way "things really are." The results showed an increase in the selection of the depictions that represented scientific representations with development, indicating that the children had acquired some scientific knowledge. Having constructed a scientific representation does not necessarily imply, however, that science learning has been completed. Students still need to learn how to distinguish scientific from phenomenal representations and manipulate them appropriately depending on the context. The results of the above-mentioned studies showed that for the majority of the astronomical phenomena investigated, many children either mixed up the two kinds of representations, thinking that scientific depictions represented "the way things appear to be" and phenomenal depictions represented "the way things really are," or selected only scientific depictions and said that they stood both for appearance and for reality. It was even more difficult for children to understand how the phenomenal depictions were related to the scientific ones and to explain why.

The development of scientific thinking skills, of which the ability to construct and manipulate multiple representations is one, has been attributed, in addition to content knowledge, to several other factors, such as information processing, executive function (working memory, shifting, and inhibition), and logical, spatial, and language abilities (e.g., Klahr, 2000; Bullock et al., 2009; Plummer, 2014; Koerber et al., 2015; Vosniadou et al., 2015; Plummer et al., 2016; Vosniadou et al., 2018). More recently, the development of scientific thinking skills has been linked to epistemological understanding (Carey et al., 1989; Carey and Smith, 1993; Kuhn, 2011), and even more recently to social cognition and theory of mind (ToM; Astington et al., 2002; Sodian et al., 2002; Osterhaus et al., 2017).

In the present research, we explore the links between children's ability to construct and manipulate scientific representations in observational astronomy, their epistemological understanding, and social cognition and more specifically, the development of a ToM. We propose (1) that children's ability to think about the differences between their beliefs and the beliefs of others in the social domain is a precursor of their ability to understand that the same event in the physical world can receive more than one interpretation and (2) that both of these abilities are related to the understanding of the constructive nature of knowledge—the understanding that our beliefs do not have an immediate relation to the world but are conjectures, hypotheses, that need to be verified and that they can be proven to be wrong. In the pages that follow, we describe the hypothesized links between ToM, personal epistemology (PE), and science learning in observational astronomy (SLOA) in greater detail.

Links Between ToM and PE

Theory of mind research investigates the development of knowledge about one's mental states such as beliefs, emotions, thoughts, and desires (Astington et al., 1988; Flavell, 2004; Apperly, 2010, 2012; Sodian and Kristen, 2016). Although children have some understanding of the basic concepts of intentionality (Meltzoff, 1995) and desire (Poulin-Dubois, 1999) from early on, they do not understand that people can have different beliefs about the same situation in the world until the age of about 4–5 years (Wimmer and Perner, 1983; Perner, 1988, 1991; Mitchell and Lacohee, 1991; Taylor et al., 1991). ToM knowledge continues to develop during the elementary school years as children come to understand that it is possible for other people to see something that they themselves cannot see and that the same object can receive different interpretations when viewed from different positions (Flavell et al., 1990; Flavell, 2000) or is seen by people with different prior knowledge (Mitchell and Lacohee, 1991; Perner, 1991; Taylor et al., 1991).

The possible links between ToM and epistemological understanding were first pointed out by researchers like Carpendale and Chandler (1996) and Kuhn et al. (2000), who argued that children's achievements in ToM mark the first step toward an increasingly interpretive view of the nature of our mental world (see also Chandler and Carpendale, 1998, and Wellman et al., 2001). Smith et al. (2000) also argued that between the ages of 4 and 6, children develop the beginnings of a PE within the framework of their ToM and then continue to reconstruct their epistemological understandings as they encounter different knowledge claims in various domains. Based on the assumption that young children's concepts are organized in intuitive theories that undergo conceptual change, Smith et al. situated children's initial epistemology as a sub-theory within their initial ToM.

Empirical support for these theoretical arguments was provided in a study by Astington et al. (2002) who investigated relations between children's performance in second-order ToM belief tasks and their epistemological understanding—namely, their ability to distinguish between the cause of a situation and a person's reason for believing it. Seventy-four children between the ages of 5 and 7 were given two- second-order false-belief (FB) tasks and two evidence tasks. In the FB tasks, the children saw the enactment of a story about a protagonist who moves a letter from place A to a new location B while mistakenly believing her friend to be absent. The children were asked to predict where the protagonist thought the friend would look for the letter. In the evidence tasks, the children also saw the enactment of two stories and were asked about the cause of an event and the character's evidence for it. For instance, in one story, a girl comes into the room and gets her feet wet without knowing that the floor was wet because a boy spilled water on it. The participants were asked why the girl's feet were wet and whether the girl knew the reason why. The results showed that performance

in the second-order FB tasks was correlated with performance in the evidence tasks and was a better predictor of them over general language and non-verbal abilities. The authors concluded that second-order FB understanding is significantly related to epistemological understanding as exhibited in children's ability to make a distinction between two epistemologically distinct entities—i.e., the cause of an event and a person's evidence for it.

Links Between ToM, PE, and Science Learning

The links between epistemological understanding and science learning were made by researchers such as Hofer and Pintrich (1997) and Kuhn (2000, 2011) early on. A training study by Sodian et al. (2002) and a recent study by Osterhaus et al. (2017) provided empirical evidence for a relation between performance in a nature of science task and experimentation skills. Other related works suggest that there are relations between epistemological understanding and a wide range of scientific thinking skills and not only experimentation (Qian and Alvermann, 1995; Mason, 2000, 2003; Koerber et al., 2015; Sodian et al., 2016).

So far, ToM has been linked to the development of experimentation skills via its relation to epistemological understanding. As mentioned earlier, ToM can be helpful in the development of a PE because it leads to the recognition that empirical data and theory are distinct entities and therefore have a different epistemological status as sources of knowledge (Astington et al., 2002; Kuhn, 2011). The development of a PE is, in turn, a precursor of the development of experimentation skills. This hypothesis was examined by Osterhaus et al. (2017) who investigated relations between advanced theory of mind (AToM), nature of science understanding, experimentation skills, and general information processing (inhibition, intelligence, and language abilities) with 402 children aged 8 to 10 years. The results indicated that AToM was an important precursor of epistemological understanding, while, in turn, children's epistemological understanding was a predictor of experimentation skills. Information processing abilities were also shown to be significantly related to experimentation skills.

The purpose of the present research was to further investigate links between advanced ToM, advanced epistemological understanding, and scientific thinking. The aspect of scientific thinking of interest in the present study was not experimentation skill but the ability to entertain dual representations of the same phenomenon in the world, and more specifically, phenomenal and scientific representations in observational astronomy. In agreement with prior research, we hypothesized that ToM would be a precursor of advanced epistemological understanding and that, in turn, PE might be a precursor of SLOA.

Furthermore, we hypothesized that there might be direct links between ToM and SLOA. Science learning requires that children can understand that the same phenomenon in the world can receive an interpretation different from that which is based on their phenomenal experience. For example, they must understand that although the Sun seems to hide behind mountains or clouds at night, the cause of the day/night cycle is to be found in the axis rotation of the Earth. They must also understand that the scientific and phenomenal interpretations do not function independently but are related to each other because they both refer to the same situation in the world, although in a different way. As mentioned earlier, the construction of a scientific representation does not mean that science learning has occurred. Only when children can both construct and reflect on different possible representations of the same phenomenon can we say that scientific learning has been achieved. As discussed earlier, the same ability underlies ToM development in the social domain. In other words, there seems to be a common cognitive/conceptual component that underlies both ToM and SLOA and that is related to the ability to construct and reflect on more than one representation of the same situation. Furthermore, the development of this ability in the social domain (ToM) appears to be a precursor of its development in the physical domain SLOA.

If there is a direct link between ToM and SLOA, what is then the role of epistemological understanding? A possible answer to this question is that epistemological understanding might create the ground—the foundation—that enables the transfer of knowledge from the social to the physical domains. ToM aids in the development of a PE because it helps children to become aware of the constructive nature of knowledge in general and the differences between theory and evidence (Sodian and Kristen, 2016). This metaconceptual understanding about the nature of knowledge, in turn, makes it more likely for children to notice that it can apply to domains other than ToM, namely, to people's explanations of phenomena in the physical world.

To sum up, we hypothesized that the emerging awareness that people can have different beliefs about events in the social world in ToM is an important precursor in children's ability to construct and reflect on different representations of the physical world SLOA. Furthermore, we hypothesized that both ToM and SLOA are conceptually linked to PE. ToM contributes to the emergence of PE as demonstrated by previous research. Developments in PE, in turn, facilitate the development of SLOA, allowing children to transfer their understanding of the possibility of alternative beliefs from social situations to the physical world. **Figure 1** demonstrates the hypothetical relation between ToM ability, PE, and SLOA.

The hypotheses were tested by examining the relation between children's performance in two ToM, one PE, and one SLOA tasks. Advanced ToM knowledge was investigated using two FB tasks in which the children had to set aside their knowledge of reality and to attribute an FB to an agent that lacked this knowledge. In the SLOA domain, children were administered a task in which they had to set aside their phenomenal perception of the physical world and adopt a scientific explanation. In both the ToM and SLOA tasks, the children were also asked to justify their responses and to reflect on what they knew and on how they knew it (Lombardi and Sinatra, 2018). The *Nature of Science Interview* (Carey et al., 1989; Smith et al., 2000; Smith and Wenk, 2006) was used to investigate epistemological understanding. We predicted that performance in all tasks will increase with age (Hypothesis 1). We also predicted that performance in the ToM tasks would be a significant predictor of performance in

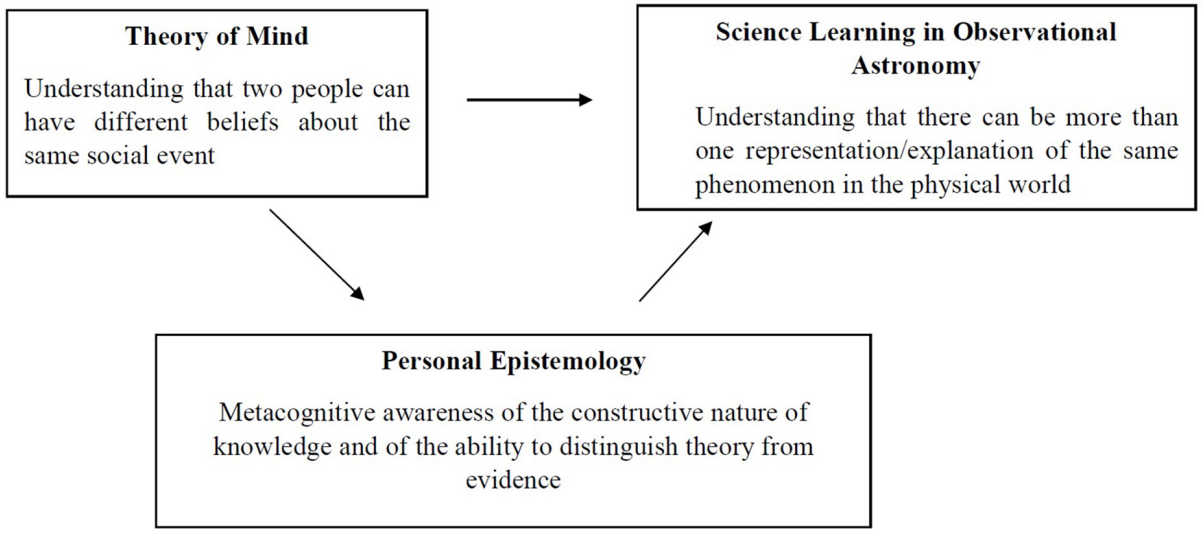


FIGURE 1 | Hypothetical relations among ToM, PE, and SLOA.

the SLOA task, even when PE and age were taken into account (Hypothesis 2). Finally, we predicted that performance in the PE task would be a significant predictor of performance in the SLOA task (Hypothesis 3).

MATERIALS AND METHODS

Participants

The participants were 63 students (34 male) who attended grades 5 and 6 in a middle-class school in central Athens. They ranged in age from 10 to 12 years and 6 months (mean age: 10 years and 7 months). There were 27 children ranging in age from 10 years to 10 years and 11 months (mean = 10 years and 3 months), 26 children ranging in age from 11 years to 11 years and 11 months (mean = 11 years and 2 months), and 10 children ranging in age from 12 years to 12 years and 6 months (mean = 12 years and 1 month).

Procedure

The students were interviewed individually at their school, in a quiet room, by one of the experimenters. The measures were administered in two sessions. The two ToM tasks and the PE task were administered first. The SLOA task was administered in the second session. Each session lasted approximately 30 to 40 min. All interviews were audio-recorded and were later transcribed for scoring.

Materials

ToM Tasks

The materials consisted of the second-order FB task "Ice Cream Story" (Perner and Wimmer, 1985) and the third-order FB task "Double Bluff" by the Strange Stories (Happé, 1994). The second-order FB task was the "Ice-Cream Story." In this story, there are two friends, John (agent A) and Mary (agent B) who want to buy ice cream. John knows that the ice cream van has been moved from a park to a church. Mary is also informed about the location change, but John does not know that Mary knows about the new location. The crucial question is where John thinks Mary will go to buy ice cream—i.e., to the park or to the church. In order to succeed in this task, the students must understand that John can have an FB about Mary's belief. In other words, that it is possible to have beliefs about other people's beliefs, and these beliefs may be false. As Perner and Wimmer (1985) discuss, there is a conflict between Mary's propositional attitude (Mary knows the van is in the church) and Mary's propositional attitude as believed by John (Mary thinks the van is in the park). There is also a conflict between what John believes about Mary's belief and about what the child knows. In this case, the child must set aside his/her own knowledge about Mary's current belief and must interpret and evaluate John's model about the ice cream van's current location.

The students were asked three questions. First, a comprehension question—"Where is the Ice-Cream Van?"—was asked in order to ensure that the story was understood. Second, a question was asked about John's belief—"Where does John think Mary has gone?"—in order to determine whether they understood that John would believe that Mary would go to the park and not to the church. Third, the children were asked the question "Why does John believe that?" to justify why they thought that John had an FB.

Third-order belief tasks investigate more advanced forms of ToM that involve feelings, motives, and the use of more complex linguistic forms such as indirect speech, irony, and white lie (Happé, 1994). In the third-order FB task, students must understand that "the intended meaning of a message is different from the literal meaning of the utterance" (Miller, 2012). The "Double Bluff" story given to the students described a situation where a soldier is interrogated by the enemy about the location of his army's tanks. In the process of the interrogation, the soldier reveals the true location of the tanks with the intention

of deceiving the enemy. In other words, the soldier thinks that the enemy will not believe that he will reveal the tanks' true location and will think that what he is telling them is not the true location of the tanks. In order to succeed in this task, students must understand that the soldier wants to deceive the enemy. In other words, the child must recognize that the soldier's utterance is intended to be interpreted non-literally. As in the previous task, the participants were also asked three questions. First a comprehension question ("Is what the prisoner said true?"), second an FB question ("Where will the enemy look for the soldier's army tanks?"), and third a justification question ("Why did the prisoner say what he said?").

PE Task

The *Nature of Science Interview* (Carey et al., 1989; Smith et al., 2000; Smith and Wenk, 2006) was used to test students' PE. This instrument examines the extent to which students have developed an understanding of the constructive nature of science. In a structured interview, the participants were asked to respond to four clusters of questions. Cluster 1 asked the students about the general aims of science and the types of questions scientists ask. Cluster 2 was about the nature and purpose of experiments and experimental procedures. Cluster 3 was about the nature of hypothesis formation and testing, and Cluster 4 was about the nature and the process of theory change. The exact questions that are used in the interview can be found in **Table 1**.

SLOA Task

The ability to reason about conflicting conceptual models of the physical world was measured using the SLOA task. The SLOA is a computer-based task that investigates students' knowledge of the scientific representations in observational astronomy and their ability to distinguish them from perceptually based representations. It also investigates the ability to reflect on the discrepancies between these two representations. The task consists of two pictures of each of six astronomical phenomena (*Earth Shape*, *Gravity*, *Relative Size of Sun and Moon*, *Relative Size of Sun and Earth*, *Day/Night Cycle*, and *Solar System*), shown in **Table 2**. One picture depicted phenomenal representations and the other scientific representations. The pictures were based on previous research in which children were asked to create their own representations of the same astronomical phenomena (Vosniadou and Brewer, 1992, 1994; Vosniadou et al., 1996). The task was validated in an earlier study with sixth-grade children (Kyriakopoulou and Vosniadou, 2008) in which the children were asked to explain what astronomical phenomena these pictures represented.

The students were shown the two pictures of each phenomenon in a random order. They were told that pictures were about the Earth, Sun, and Moon and the specific astronomical phenomenon investigated (e.g., the shape of the Earth, where people live on the Earth, the relative size of the Earth, and the Moon, etc.). They were then asked the following questions: (1) "Look at the two pictures. What does the first picture show? What does the second picture show? These questions tested children's understanding of the referential nature of these pictures. They also revealed if they had been exposed to a scientific explanation of them." (2) What are the differences between these two pictures? This question tested whether the children were able to understand that the two pictures referred to the same phenomenon in the world. If the students did not refer to the distinction between appearance and reality at the time of the second question, the experimenter prompted them to select which picture was closer to the way

TABLE 1 | The Nature of Science Interview (Carey et al., 1989; Smith et al., 2000; Smith and Wenk, 2006).

THE NATURE OF SCIENCE INTERVIEW

1. General Aims of Science—Types of Questions

1.1 Our friend John has some questions. He read some things about what science is about and what scientists do and wants our help to understand what he read. He also wants to hear what you think about science and scientists. What do you think the word "scientist" means? Can you give him an example?

1.2 What sorts of things do scientists do? How do they reach their goals?

1.3 Do scientists ask questions? Can you give me a specific example of a question that a scientist would ask?

1.4 What would scientists do to answer their question?

2. Nature and Purpose of Experiments

2.1 What is an experiment?

2.2 Do scientists do experiments? Why do scientists do them? In general, how do scientists decide what experiment to do?

3. Nature of Hypotheses

3.1 Many times, we make hypotheses about various things. Have you ever heard the word hypothesis? What do you think is a hypothesis a scientist does? Do you think a hypothesis is the same as a guess or do you think that there is a difference? What is the difference?

3.2 Can you give a specific example of a scientific hypothesis?

4. Nature of Theories—Process of Theory Change

4.1 Do scientists have ideas/theories about the world?

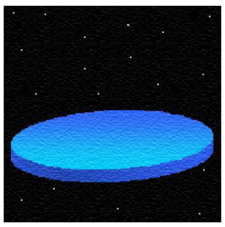
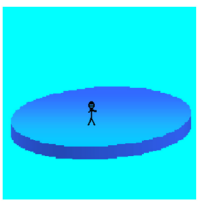
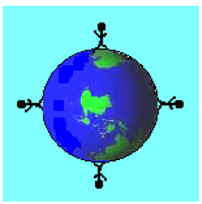
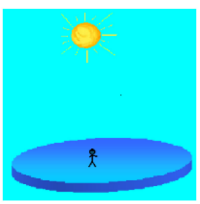
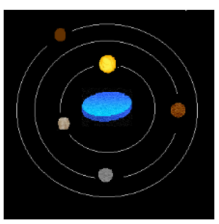
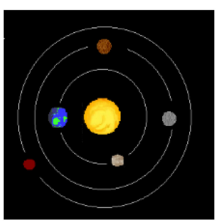
4.2 What is a scientific theory? Can you give a specific example of a scientific theory?

4.3 Do you think a scientist's ideas influence the way he tries to find answers to his questions?

4.4 Say two scientists believe different things about our world. How can we decide which one is right?

4.5 Do scientists ever change their hypotheses or theories? When would they do that and why? Can scientists make mistakes or be wrong? How? Do scientists always achieve their goals? Why?

TABLE 2 | Science Learning in Observational Astronomy Task.

	SLOA task		
Astronomical phenomenon	***Question 1*: Look at the two pictures. What does the first picture show? What does the second picture show?** ***Question 2*: What are the differences between these two pictures?**		***Question 3*: Justification question**
Earth's shape			How do you explain that the Earth seems flat, when at the same time we accept that it is a sphere?
Where people live on the earth (Gravity)			Since people can stay everywhere on Earth, can you tell me where this ball would fall if someone was here at the bottom of the Earth and here at the top of the Earth? (the experimenter shows where) Why is this happening?
Sun–Moon relative size			Why do the Sun and the Moon look like they have the same size, when in fact the Sun is bigger than the Moon?
Sun–Earth relative size			Why does the Sun seem smaller to us, when in fact it is much larger than the Earth?
Solar system			Why do we not understand the Earth's movement around the Sun? Why do you say that the Earth moves when we do not feel its movement?
Day–Night cycle			Why do we see the Sun rising from the East and setting from the west, when we know that it is the Earth that moves and not the Sun?

things appear to be and which picture showed the way things are in reality, clarifying that they could choose the same picture for both questions. Question 3—the justification question—was asked only if the students referred to the appearance–reality distinction in the second question. The justification questions for each phenomenon are shown in **Table 2**.

TABLE 3 | Type of student responses in each question cluster in the Nature of Science Interview by epistemic level.

Cluster of questions	Epistemic level				
	Level 1	**Level 1.5**	**Level 2**	**Level 2.5**	**Level 3**
Cluster 1: Q1.1–1.4 General aims of science and type of scientists' questions	Scientists simply find or discover new information and ask procedural and journalistic questions	Scientists try to find out how something works (unclear if they refer to a procedure or a mechanism)	Scientists try to find underlying mechanisms, ask questions about theoretical entities and reflective questions about their ideas	Scientists formulate questions to find out how something works	Discuss how multiple levels of questions interact
Cluster 2: Q2.1–2.2 Nature and purpose of experiments and experimental procedures	Experiment is similar to producing a desirable outcome	Experiment involves measuring variables/scientists do experiments to find out how something works	Scientists test their ideas by an experiment	Experiments are a way to test competing hypotheses	Experiments test causal ideas
Cluster 3: Q3.1–3.2 Nature of hypothesis formation and theory testing	No differentiation between hypotheses and experiments	Hypotheses are similar to guesses	Scientists understand and evaluate their own ideas through experimentation	Hypotheses are defined as explanatory ideas	Hypothesis testing provides evidence for/against a theory
Cluster 4: Q4.1–4.7 Nature and process of theory change	Scientists may abandon or change an idea based on a single experiment or observation	Hypotheses can change, but theories do not, and change occurs by doing experiments	Scientists develop new ideas	The ideas scientists investigate are more complex, and it takes work to understand them	Ideas are embedded in theoretical frameworks that constrain the generation of new hypotheses

Scoring

ToM Tasks

Responses to ToM tasks (Perner and Wimmer, 1985; Happé, 1994) were placed in five categories based on the recognition of FB and the type of justification provided in both tasks. The response categories were the following: no recognition of FB (score 1); recognition of second-order FB only (score 1.5); recognition of second-order FB and correct justification (score 2); recognition of both second- and third-order FBs and correct justification for the second-order belief (score 2.5); and recognition of second- and third-order FBs and correct justification for both (score 3). Two independent coders placed the students in one of the above-mentioned categories. Agreement rate was calculated to be 97% and was statistically significant based on Kendall's tau correlation analysis ($\tau = 0.965$, $N = 63$; $p < 0.001$). All disagreements were resolved through discussion. Two students recognized both the second- and third-order FBs but provided correct justification only for the third-order FB task, while three students recognized and correctly justified only the third-order FB. These students were placed in the more advanced category 5 because they succeeded in the most advanced task. The reliability of children's responses in the second-order FB task and third-order FB task was not very high (Cronbach's $\alpha = 0.43$). This was probably due to the small number of items (two tasks—second and third order—and two responses per task, correct vs. incorrect recognition, and justification or not). Prior studies in which these tasks were used have shown that they were valid assessments of children's ToM ability.

PE Task

The *Nature of Science Interview* task (Carey et al., 1989; Smith et al., 2000; Smith and Wenk, 2006) was scored based on the system developed by Carey and her colleagues (Carey et al., 1989; Carey and Smith, 1993; Smith et al., 2000; Smith and Wenk, 2006), consisting of four question clusters, each of which had five epistemic levels. **Table 3** shows the type of student responses for each question cluster by epistemic level.

Each epistemic level was scored as 1, 1.5, 2, 2.5, and 3, respectively. Two coders independently scored all of students' responses. Agreement rate was calculated to be 98% for Cluster 1 ($\tau = 0.976$, $N = 63$; $p < 0.001$), 98% for Cluster 2 ($\tau = 0.985$, $N = 63$; $p < 0.001$), 94% for Cluster 3 ($\tau = 0.940$, $N = 63$; $p < 0.001$), and 92% for Cluster 4 ($\tau = 0.921$, $N = 63$; $p < 0.001$). All disagreements were resolved through discussion. Reliability for children's responses for all the 13 questions was Cronbach's $\alpha = 0.70$.

Each student was given one score for each cluster, based on the highest epistemic level achieved in the cluster questions. A final average epistemic level score for each student was also computed based on the students' level scores in the four clusters. The final epistemic level scores represented the following competencies: epistemic level 1 responses (*knowledge unproblematic epistemology*) agree with the belief in true and certain knowledge. Students refer to scientists' ideas, experiments, and results in an undifferentiated mode, and goals are the activities and products of science. Epistemic level 1.5 responses are more elaborated concepts of level 1.

Students become increasingly aware that scientists have ideas but do not yet understand that these ideas are tested through experimentation. Epistemic level 2 responses show sensitivity to explanation and hypothesis testing. Students begin to differentiate between ideas, experiments, and results. Epistemic level 2.5 responses are more complex and sophisticated expressions of level 2. The students at this level show a more complex understanding of the testing process with reference to multiple pieces of evidence and begin to understand that the development of ideas is not just a process of simply adding new ideas to preexisting ones. Epistemic level 3 responses (*knowledge problematic epistemology*) reveal an understanding of the uncertain and relative nature of knowledge (for a detailed description, see also Carey and Smith, 1993; Smith et al., 2000).

SLOA Task

A fourth-step process was followed to score the students' responses with the SLOA task. First, for each astronomical phenomenon, each student selected one picture for appearance and one for reality. Based on their choices, the students were placed in one of the following four categories: (1) no distinction between appearance and reality—when the picture chosen was the phenomenal one for both reality and appearance; (2) appearance–reality reversed—when the picture chosen was the phenomenal one for reality and the scientific one for appearance; (3) scientific responses only—when the picture chosen was the scientific one for both reality and appearance; and (4) distinction—when the picture chosen was the scientific one for reality and the phenomenal one for appearance. Two coders independently scored all responses. There were no disagreements between the two independent coders ($\tau = 1$, $n = 63$; $p < 0.001$).

Second, for each astronomical phenomenon, the students were asked to justify their choices. Depending on the type of the justification provided, their responses were placed in the following three categories: (1) no justification, if they could not make the appearance–reality distinction; (2) initial, if they were consistent with phenomenal experience or if they revealed any kind of phenomenal misunderstanding; and (3) scientific, if they could justify the differences between appearance and reality in scientific terms. Kendall's tau correlation analysis showed that the agreement between the coders was statistically significant ($\tau = 0.977$, $n = 63$; $p < 0.001$). All disagreements were resolved through discussion.

Third, based on the pictures selected and on their justifications, the children were placed in one of the five overall SLOA categories and were given the overall scores 1, 1.5, 2, 2.5, and 3, respectively, for each astronomical phenomenon, as shown in **Table 4**. In this step, there was total agreement between the coders in how they applied the five-level coding system ($\tau = 1$, $n = 63$; $p < 0.001$).

Fourth, a final score for each student was also calculated based on the mean score in the combined six astronomical phenomena, and this final score was used in the statistical analyses. The reliability of children's final SLOA responses for the six astronomical phenomena was $\alpha = 0.76$.

RESULTS

Descriptive Statistics

Table 5 shows the frequency and percentage of students assigned in the final response categories in the ToM task as a function of age.

Table 6 shows the frequency and percentage of students assigned to each epistemic level in the four question clusters. As can be seen, the majority of the students were categorized at epistemic level 1. When looking at the performance of the students as a function of age, we can see (**Table 7**) an increase in epistemic level as predicted, particularly for the older students in the sample.

Table 8 shows the frequency and percentage of students placed in each of the five response categories as a function of age. As predicted, performance increased with age, particularly in the number of children who could provide justifications of the distinction between phenomenal and scientific depictions.

Correlations

Children's final scores in the three tasks were considered to represent meaningful and equally spaced intervals indicating progressively higher levels of response. Pearson correlations showed significant results between age and performance in all the tasks: ToM tasks ($r = 255$, $p < 0.05$), PE ($r = 468$, $p < 0.01$),

TABLE 4 | Presentation of the scoring process for each astronomical phenomenon in the SLOA Task.

First step			Second step	Third step	
Appearance–Reality distinction			Type of justification	Final total categories	
	Pictures chosen				
Response categories	Appearance	Reality	Response categories	Response categories	Overall score
No distinction	Phenomenal	Phenomenal	No justification	No distinction/no justification	1
A–R reversed	Scientific	Phenomenal	No justification	A–R reversed/no justification	1.5
Scientific responses	Scientific	Scientific	No justification	Scientific responses/no justification	2
Distinction	Phenomenal	Scientific	Initial justification	Distinction/initial justification	2.5
			Scientific justification	Distinction/scientific justification	3

TABLE 5 | Frequency and percentage of students in the five categories of the two ToM tasks combined as a function of age (Perner and Wimmer, 1985; Happé, 1994; *N* = 63).

Response categories for ToM tasks		Age variation			
		10 *N* = 27	**11 *N* = 26**	**12 *N* = 10**	***Total N = 63***
1	No recognition of false belief	5 (19%)	5 (19%)	1 (10%)	*11 (18%)*
1.5	Recognition of second-order false belief only	7 (26%)	5 (19%)	–	*12 (19%)*
2	Recognition of second-order false belief and correct justification	9 (33%)	3 (12%)	4 (40%)	*16 (25%)*
2.5	Recognition of both second-order and third-order false belief and correct justification for the second-order belief	–	1 (4%)	1 (10%)	*2 (3%)*
3	Recognition of second- and third-order false belief and correct justification for both	6 (22%)	12 (46%)	4 (40%)	*22 (35%)*

TABLE 6 | Frequency and percentage of students in the four clusters of questions as a function of epistemic level at the *Nature of Science Interview* (Carey et al., 1989; Smith et al., 2000; Smith and Wenk, 2006; *N* = 63).

Epistemic level in the nature of science interview	Clusters of questions			
	Cluster 1: Q1.1–1.4 General aims of science and type of scientists' questions	**Cluster 2: Q2.1–2.2 Nature and purpose of experiments and experimental procedures**	**Cluster 3: Q3.1–3.2 Nature of hypothesis formation and theory testing**	**Cluster 4: Q4.1–4.7 Nature and process of theory change**
1 Level 1: Knowledge unproblematic epistemology	52 (83%)	52 (83%)	52 (83%)	42 (67%)
1.5 Elaborated Level 1	4 (6%)	2 (3%)	10 (16%)	20 (32%)
2 Level 2: Transitional ideas: Introduction of explanation and hypothesis testing	7 (11%)	9 (14%)	1 (2%)	1 (2%)
2.5 Elaborated Level 2	–	–	–	–
3 Level 3: Knowledge problematic epistemology	–	–	–	–

TABLE 7 | Frequency of students in each epistemic level based on their total responses in the Nature of Science Interview as a function of age (*N* = 63).

Epistemic level in the nature of science interview		Age variation			
		10 *N* = 27	**11 *N* = 26**	**12 *N* = 10**	***Total N = 63***
1	Level 1: Knowledge unproblematic epistemology	26 (96%)	21 (81%)	5 (50%)	*52 (82%)*
1.5	Elaborated Level 1	1 (4%)	3 (11%)	2 (20%)	*6 (10%)*
2	Level 2: Transitional ideas: Introduction of explanation and hypothesis testing	–	2 (8%)	3 (30%)	*5 (8%)*
2.5	Elaborated Level 2	–	–	–	–
3	Level 3: Knowledge problematic epistemology	–	–	–	–

TABLE 8 | Frequencies and percentage of students in the five categories based on their total responses in the SLOA Task as a function of age (*N* = 63).

Response categories for observational astronomy task		Age variation			
		10 *N* = 27	**11 *N* = 26**	**12 *N* = 10**	***Total N = 63***
1	No Appearance–Reality distinction/No justification	4 (15%)	–	–	*4 (6%)*
1.5	Appearance–Reality reversed/No justification	2 (7%)	2 (8%)	–	*4 (6%)*
2	Scientific responses only/No justification	7 (26%)	5 (19%)	1 (10%)	*13 (21%)*
2.5	Distinction/Initial Justification	14 (52%)	14 (54%)	5 (50%)	*33 (52%)*
3	Distinction/Scientific Justification	–	5 (19%)	4 (40%)	*9 (14%)*
	Total	*27*	*26*	*10*	*63*

and SLOA task (r = 440, $p < 0.01$). **Table 9** shows Pearson product–moment correlations between age (range of age: 10 years to 12 years and 6 months) and scores on ToM tasks (M = 2.09, SD = 0.76), Nature of Science Interview (PE; M = 1.14, SD = 0.20), and SLOA task (SL; M = 2.23, SD = 0.39). As hypothesized, performance in ToM, PE, and SLOA correlated significantly with each other, and all were correlated with age. Because of the skewness in the PE variable, the correlations were also

TABLE 9 | Pearson correlations of age, ToM, PE, and SLOA (N = 63).

	Age	ToM	PE	SLOA
Age	–			
ToM	0.255*	–		
PE	0.468**	0.431**	–	
SLOA	0.440**	0.474**	0.531**	–

*$p \leq 0.05$. **$p \leq 0.01$.

conducted using the logarithmic transformation of the PE. The results did not change.

Regressions

Two hierarchical multiple regression analyses were conducted in order to examine the following hypotheses: (1) that both ToM and PE predict students' performance in the SLOA task and (2) that both independent variables (ToM and PE) will continue to be good predictors for the dependent variable (SLOA) even when age was taken into consideration.

A two-step hierarchical regression analysis examined the first hypothesis. In this analysis, performance in the SLOA task was the dependent variable. We first introduced into the equation performance in the ToM tasks as a predictor. At the second step, performance in PE task was added. The results showed that at Step 1, ToM ability contributed significantly to the regression model [$F(1, 62) = 17.707$, $p \leq 0.001$], and accounted for 23% of the variation in the SLOA task. The introduction of the PE variable explained an additional 13% of the variation in performance in the SLOA task [$F(2, 62) = 16.558$, $p \leq 0.001$]. In the final equation, PE made a greater contribution ($\beta = 0.401$, $p \leq 0.001$) than ToM ($\beta = 0.302$, $p \leq 0.001$) (see **Table 10**).

A three-step hierarchical multiple regression analysis was conducted in order to examine the second hypothesis. In this analysis, performance on the SLOA task was the dependent variable. We first introduced into the equation the variable *age* as predictor. At the second step, performance in the ToM tasks was added. At the third step, performance in PE task was added in addition to age and ToM. A significant regression equation was found in all three steps (see **Table 11**). In Step 1, age was found to contribute significantly to the prediction of performance on the SLOA task [$F(1, 62) = 14.632$, $p \leq 0.001$), explaining 19% of the variance, In Step 2, the introduction of ToM ability explained an additional 14% of the variance [$F(2, 62) = 15.020$, $p \leq 0.001$]. The inclusion of PE in Step 3 also produced a significant regression equation [$F(3, 62) = 12.851$, $p \leq 0.001$], explaining an additional 6% of the variance. As we see in **Table 11**, in the final equation, PE made a greater contribution ($\beta = 0.301$, $p \leq 0.01$) than ToM ($\beta = 0.287$, $p \leq 0.01$) whereas age made the least contribution ($\beta = 0.226$, NS).

Both regressions were repeated using the logarithmic value for the PE variable. The results did not change.

A path analysis using the Analysis of Moment Structure (AMOS 26) was used to test the hypothesized direct effects from ToM to SLOA and the mediating role of PE. The analysis resulted in a saturated, just-identified model ($\chi^2 = 0$, $df = 0$; GFI = 1, CFI = 1, and NFI = 1), indicating a perfect fit. All path coefficients were statistically significant, indicating positive direct effects from ToM on SLOA ($\beta = 302$, $p \leq 0.01$), from PE on SLOA ($\beta = 401$, $p \leq 0.001$), and from ToM on PE ($\beta = 431$, $p \leq 0.001$) (see **Figure 2**). The indirect effect of ToM on SLOA through PE was also statistically significant ($p \leq 0.01$, 95% CI [0.074, 0.296]).

DISCUSSION

Previous research has examined relations between social cognition, epistemological understanding, and scientific

TABLE 10 | Summary of hierarchical regression analysis predicting performance on the SLOA Task from ToM and PE.

Step/variable added	*B*	SEB	β	*R*	R^2	Δ*R*	Δ*F*
Step 1: ToM	0.249	0.059	0.474**	0.474	0.225		
Step 2				0.596	0.356	0.131	12.167**
ToM	0.159	0.060	0.302*				
PE	0.799	0.229	0.401*				

*$p \leq 0.01$. **$p \leq 0.001$.

TABLE 11 | Summary of hierarchical regression analysis predicting performance on the SLOA Task from Age, ToM and PE.

Step/variable added	*B*	SEB	β	*R*	R^2	Δ*R*	Δ*F*
Step 1: Age	0.252	0.066	0.440***	0.440	0.193		
Step 2				0.578	0.334	0.140	12.621**
Age	0.195	0.062	0.341**				
ToM	0.203	0.057	0.387**				
Step 3				0.629	0.395	0.062	6.007*
Age	0.129	0.066	0.226				
ToM	0.151	0.059	0.287**				
PE	0.601	0.245	0.301**				

*$p \leq 0.05$. **$p \leq 0.01$. ***$p \leq 0.001$.

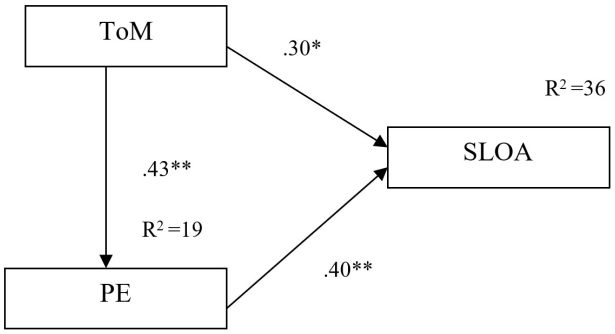


FIGURE 2 | Path model showing a direct path from ToM to PE and from PE to SLOA, as well as a direct path from ToM to SLOA.

thinking, focusing on experimentation. The focus of the present research was on a different aspect of scientific thinking that has to do with the ability to construct and reflect on more than one representation of the same phenomenon in the physical world. Many researchers have discussed the importance of the ability not only to construct but also to move accurately and flexibly among various levels of representations for science learning and conceptual change (Spada, 1994; Pozo et al., 1999; Kozma, 2003; Wiser and Smith, 2009). Recent research regarding the coexistence of phenomenal and scientific representations (Shtulman and Valcarcel, 2012) and the interference of phenomenal representations in scientific reasoning tasks with adults (Vosniadou et al., 2018) further underscores the importance of developing students' ability to think with more than one representation. In the present study, we examined children's ability to construct and reflect on phenomenal and scientific representations in observational astronomy, a domain of science whose development is well understood based on prior research.

Science learning in observational astronomy requires that children construct explanations of phenomena such as the shape of the Earth and the day/night cycle, which are different from those suggested by perceptual experience. It also requires that they understand how this scientific representation is related to their phenomenal experience. We argued that there are some similarities between this type of science learning and correct performance in second- and third-order ToM tasks. ToM tasks are usually based on the understanding that two people can have different beliefs about the same social event or that the same proposition may have two meanings, one literal and one non-literal, depending on the context (see also Moll and Tomasello, 2007; Tomasello and Moll, 2010). In other words, ToM and science learning seem to share a common underlying cognitive/conceptual component, having to do with the ability to construct and reflect upon two representations of the same situation. It was hypothesized that this ability will develop with age and that its development in the social domain would be a precursor of its development in science learning.

The results of the present research showed age-related increases in performance in both the ToM and the science learning tasks (Hypothesis 1). Regarding the ToM tasks, the findings revealed a significant correlation between ToM and age. The younger children (10-year-olds) were able to answer correctly the second-order belief task but found it difficult to justify their responses. Third-order FB understanding with correct justification increased in the 12- to 12.5-year-old group. These results are in accordance with existing research (Happé, 1994; Muris et al., 1999; White et al., 2009). Children's developing ability to reflect on their reasoning (Miller, 2012) has been shown to be related to the ability to understand and imagine multiple perspectives and alternatives (Bosacki and Astington, 1999) and to epistemological thinking (Mason, 2002).

Regarding the astronomy task, the results also showed significant improvements with age. Sixty-six percent of the students distinguished the depiction that best represented "the way things appear to be" from "the way things really are" across the various astronomical phenomena (Categories 4 and 5), indicating that they knew the scientific explanation. However, the ability to verbally articulate the relation between phenomenal and scientific depictions increased, particularly in the 12- to 12.5-year age group. The results agree with previous research (Vosniadou and Kyriakopoulou, 2006).

Successful performance in the astronomy task undeniably requires domain-specific content knowledge as well as the development of complex perspective taking, spatial reasoning, information processing, and executive function skills. However, it also critically depends on the ability to understand that it is possible to entertain more than one representation of the same phenomenon in the physical world and reflect on these representations. The results of the regression confirmed that performance in the ToM tasks was a significant predictor of performance in the astronomy task. The present study is the first to show that ToM is a precursor of science learning. Previous research showed ToM to be a precursor of epistemological understanding only, but not of scientific thinking. This result supports the hypothesis that there is a common conceptual component between ToM and SLOA. Both tasks require the ability to construct and reflect upon dual representations. It appears that this ability develops first in the social domain. Understanding that people have different beliefs about the same social event facilitates the recognition that it is possible that the same phenomenon in the physical world might receive different interpretations.

Students' ideas about the nature of science (Carey et al., 1989; Smith et al., 2000; Smith and Wenk, 2006) were found to be mostly in agreement with an initial level of PE (level 1), indicating an attachment to a single, true, and certain truth. These results are consistent with the findings of Carey et al. (1989); Honda (1994), and Smith et al. (2000) who used the same measure to investigate the epistemic beliefs of students attending the last grades of primary school and first grades of secondary school. In the present research, only a few of the older students (12–12.5 years old) gave responses that were categorized in level 2. It seems that for these few children, there is a small, although significant, shift from an entirely

objective and certain view of knowledge to a more constructivist epistemological stance, where there is a need to think explicitly about their beliefs, examine them in a framework of alternatives, and provide the evidence to confirm/disconfirm them (Kuhn, 1993). The development of these dispositions to think about knowledge, the nature of science, and the process of knowing in a framework of conflicting views may serve as the foundation for the development of scientific thinking (Qian and Alvermann, 1995; Mason, 2000). Indeed, the results of the regression analysis confirmed the hypothesis regarding the relations between ToM, science learning, and epistemological understanding, by showing that performance in the nature of science task was a significant predictor of performance in the astronomy task, even when ToM and age were taken into consideration, explaining the largest percentage of the variance (Hypothesis 3).

A possible interpretation of the role of PE is that it acts as a mediator between ToM and SLOA, allowing the transfer of knowledge from the social to the physical domains. In other words, the recognition, first achieved in the ToM domain, that people can have different beliefs or that the same event may be interpreted in different ways becomes the foundation that allows children to form an awareness of the constructive nature knowledge in general (Astington et al., 2002; Eisbach, 2004; Sodian and Kristen, 2016; Osterhaus et al., 2017). It can also serve as the foundation for the recognition that evidence and theory are distinct entities, an understanding that is central to scientific learning (Kuhn et al., 2000; Iordanou, 2016). The development of a PE, in turn, further enables and facilitates the transfer of ToM understanding in the domain of science. Although a preliminary investigation of the possible mediating role of PE between ToM and SLOA was confirmed, this relation needs to be investigated in greater detail in future research with a larger number of participants and taking into effect especially the role of age and of executive function skills.

The awareness that different people can have different interpretations of the same social event and the ability to verbally articulate it and generalize it to other knowledge situations is not a trivial task. It constitutes a major cognitive development and requires considerable conceptual changes, similar to that described by Karmiloff-Smith (1992) in the area of language development. Karmiloff-Smith investigated how children treat the fact that the same word may have more than one meaning, i.e., that it can refer to two different situations in the world. Although in our case children face a reverse problem (where two different representations refer to the same situation in the world), nevertheless, in both cases, we have a problem that demands dealing with dual representations of some form. Karmiloff-Smith's studies (1979) showed that 5- to 6-year-olds were not always able to use the words correctly and explain why. Only later, around the age of 6–7 years, were they able to do so, suggesting that they had achieved a consciously accessible and verbally stated metalinguistic knowledge. Karmiloff-Smith (1992) argued that this achievement was the product of a process of "representational re-description." Through this process, implicit information becomes explicit knowledge, progressively available to other parts of the cognitive system and under self-evaluation. Then, it is feasible for the child to produce and use multiple representations at different levels of explicitness and detail.

In more recent work, Kuhn (2006) relates similar achievements to developments in executive control in adolescence that allow metacognitive reflection of one's representations and flexible access to dual representations. According to Kuhn, the absence of this ability leads to a singular experience of "the way things are" as a framework for understanding the world.

The results of the present research indicate that level 2 epistemological understanding was only beginning to be achieved in the 12-year-old group. As the regression analysis showed, epistemological understanding was the most important predictor of the participants' ability to reflect on conflicting astronomical depictions. The findings support the interpretation of the possible mediating role of PE and a conceptual link of ToM, science learning, and epistemological understanding. As children moved to a more constructivist epistemology of science, they succeeded more often in forming scientific representations and understanding their relation to their phenomenal experience.

Limitations and Future Directions

The present research provides us with an initial understanding of the links between social cognition, epistemological understanding, and science learning in the area of observational astronomy but has several limitations. One limitation is the small sample size and the limited statistical analysis. The research needs to be replicated with a larger developmental sample that would allow the use of structural equation modeling to further test the presence of direct links between ToM and science learning, the hypothesized mediating role of epistemological knowledge, and the roles of age and executive function skills. In addition, other important factors such as spatial and perspective-taking abilities, language abilities, and prior knowledge should be considered in future research. It is very probable that particularly executive function skills, such as working memory, inhibition, and shifting, could account for part of the commonalities observed between ToM, PE, and SLOA. Future research needs also to further test the links between ToM and science learning in other domains of science, where there is reason to believe that common conceptual components are shared, not only in the domain of observational astronomy.

It is possible that the relations between social cognition, epistemological understanding, and science learning are bidirectional and not unidirectional. For example, it is possible that developments in students' PE allow for a more advanced understanding of ToM, while the learning of science makes possible the development of more sophisticated epistemological understandings. This hypothesis needs to be tested in future research.

Last but not least, it is important to test the relations between ToM, epistemological understanding, and science learning using training and intervention experiments. If ToM and PE are precursors of science learning, we should expect that training in ToM, and/or ToM and PE at an early age improves

children's ability to construct and flexibly manipulate different representations of the same phenomenon in the physical world.

Implications for Science Instruction

Many researchers have argued that science education should be oriented toward the development of students' ability to construct multiple representations and be able to move flexibly among them (Spada, 1994; Pozo et al., 1999; Wiser and Amin, 2001; Treagust et al., 2017). According to Schwartz and Brown (2013), improvements in students' ability to move across various levels of representation (e.g., phenomenal, molecular, and symbolic) could lead to greater scientific understanding. Children's performance in the present observational astronomy task indicates that children, even when they have knowledge of the scientific representations, still find it difficult to navigate between them and the phenomenal ones and reflect upon them. Other research have also shown that students' ability to integrate the use of multiple representations during science learning is limited. Students tend to focus on surface features and ignore underlying mechanisms and/or are unable to coordinate between different representations (Kozma, 2003; Seufert, 2003; Won et al., 2014).

It has been difficult to develop science instruction that can improve students' ability to form flexible scientific representations and understand their relation to perceptual experience. The importance of information processing and executive function limitations, such as working memory, inhibition, shifting, and spatial reasoning, has often been emphasized. There is no doubt that these are very important. There have been few suggestions, however, as to how to strengthen children's conceptual understanding. The present results suggest that instruction that focuses on ToM and PE might help children develop the conceptual understanding necessary to grasp the constructive nature of knowledge and the distinction between theory and evidence, paving the way for improved learning in science.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are not publicly available. Requests to access the datasets should be directed to the corresponding author.

ETHICS STATEMENT

The study was reviewed and approved by the Institute of Educational Policy of the Greek Ministry of Education and Religious Affairs. Written informed consent to participate in this study was provided by the participants' legal guardian.

AUTHOR CONTRIBUTIONS

NK and SV contributed to the conception and design of the study. NK acquired the data and organized the database. NK and SV analyzed and interpreted the data and wrote sections of the manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

The research was supported with funds from Professor Vosniadou's special research account with the National and Kapodistrian University of Athens, Research Project K.E. 8851.

ACKNOWLEDGMENTS

We would like to thank all participating children and their teachers and directors.

REFERENCES

Apperly, I. (2010). *Mindreaders: the Cognitive Basis Of Theory Of Mind*. London: Psychology Press.

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Q. J. Exper. Psychol.* 65, 825–839. doi: 10.1080/17470218.2012.676055

Astington, J. W., Harris, P. L., and Olson, D. R. (1988). *Developing Theories of Mind*. Cambridge: Cambridge University Press.

Astington, J. W., Pelletier, J., and Homer, B. (2002). Theory of mind and epistemological development: the relation between children's second-order false-belief understanding and their ability to reason about evidence. *New Ideas Psychol.* 20, 131–144. doi: 10.1016/s0732-118x(02)00005-3

Bosacki, S., and Astington, J. W. (1999). Theory of mind in preadolescence: relations between social understanding and social competence. *Soc. Dev.* 8, 237–255. doi: 10.1111/1467-9507.00093

Bullock, M., Sodian, B., and Koerber, S. (2009). "Doing experiments and understanding science: development of scientific reasoning from childhood to adulthood," in *Human Development From Early Childhood to Early Adulthood. Findings From The Munich Longitudinal Study*, eds W. Schneider and M. Bullock (Mahwah, NJ: Erlbaum), 173–197.

Carey, S. (1985). *Conceptual Change In Childhood*. Cambridge, MA: MIT Press.

Carey, S., Evans, R., Honda, M., Jay, E., and Unger, C. (1989). '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. *Intern. J. Sci. Educ.* 11, 514–529. doi: 10.1080/0950069890110504

Carey, S., and Smith, C. (1993). On understanding the nature of scientific knowledge. *Educ. Psychol.* 28, 235–251.

Carpendale, J., and Chandler, M. (1996). On the distinction between false belief understanding and subscribing to an interpretive theory of mind. *Child Dev.* 67, 1686–1706.

Chandler, M. J., and Carpendale, J. I. (1998). "Inching toward a mature theory of mind," in *Self-Awareness: Its Nature And Development*, eds M. D. Ferrari and R. J. Sternberg (New York, NY: Guilford Press), 148–190.

Chi, M. T. (2009). "Three types of conceptual change: belief revision, mental model transformation, and categorical shift," in *International Handbook Of Research On Conceptual Change*, ed. S. Vosniadou (Abingdon: Routledge), 89–110.

DeWolf, M., and Vosniadou, S. (2015). The representation of fraction magnitudes and the whole number bias reconsidered. *Learn. Instr.* 37, 39-49. doi: 10.1016/j.learninstruc.2014.07.002

Eisbach, A. O. D. (2004). Children's developing awareness of diversity in people's trains of thought. *Child Dev.* 75, 1694–1707. doi: 10.1111/j.1467-8624.2004.00810.x

Flavell, J. H. (2000). Development of children's knowledge about the mental world. *Intern. J. Behav. Dev.* 24, 15–23. doi: 10.1080/016502500383421

Flavell, J. H. (2004). Theory-of-mind development: retrospect and prospect. *Merrill Palmer Q.* 50, 274–290. doi: 10.1353/mpq.2004.0018

Flavell, J. H., Green, F. L., and Flavell, E. R. (1990). Developmental changes in young children's knowledge about the mind. *Cogn. Dev.* 5, 1–27. doi: 10.1016/0885-2014(90)90010-q

Happé, F. G. E. (1994). An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *J. Autism Dev. Disord.* 24, 129–154. doi: 10.1007/bf02172093

Hofer, B. K., and Pintrich, P. R. (1997). The development of epistemological theories: beliefs about knowledge and knowing and their relation to learning. *Rev. Educ. Res.* 67, 88–140. doi: 10.3102/00346543067001088

Honda, M. (1994). *Linguistic Inquiry In The Science Classroom: "It Is Science, But It's Not Like A Science Problem In A Book*. Cambridge, MA: MIT Press.

Iordanou, K. (2016). From theory of mind to epistemic cognition. a lifespan perspective. *Front. Learn. Res.* 4, 106–119. doi: 10.14786/flr.v4i5.252

Karmiloff-Smith, A. (1979). Micro-and macrodevelopmental changes in language acquisition and other representational systems. *Cogn. Sci.* 3, 91–117. doi: 10.1016/s0364-0213(79)80027-0

Karmiloff-Smith, A. (1992). *Beyond Modularity*. Cambridge, MA: MIT Press.

Klahr, D. (2000). *Exploring Science: The Cognition And Development OfDiscovery Processes*. Cambridge, MA: MIT Press.

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., and Sodian, B. (2015). The development of scientific thinking in elementary school: a comprehensive inventory. *Child Dev.* 86, 327–336. doi: 10.1111/cdev.12298

Kozma, R. B. (2003). Technology and classroom practices: an international study. *J. Res. Technol. Educ.* 36, 1–14. doi: 10.1080/15391523.2003.10782399

Kuhn, D. (1993). Science as argument: implications for teaching and learning scientific thinking. *Sci. Educ.* 77, 319–337. doi: 10.1002/sce.3730770306

Kuhn, D. (2000). Metacognitive development. *Curr. Direct. Psychol. Sci.* 9, 178–181.

Kuhn, D. (2006). Do cognitive changes accompany developments in the adolescent brain? *Perspect. Psychol. Sci.* 1, 59–67. doi: 10.1111/j.1745-6924.2006.t01-2-.x

Kuhn, D. (2011). "What is scientific thinking and how does it develop?," in *Handbook of Childhood Cognitive Development*, 2nd Edn, ed. U. Goswami (Oxford: Blackwell), 472–523.

Kuhn, D., Cheney, R., and Weinstock, M. (2000). The development of epistemological understanding. *Cogn. Dev.* 15, 309–328. doi: 10.1016/s0885-2014(00)00030-7

Kyriakopoulou, N., and Vosniadou, S. (2004). Ontological and epistemological problems in the distinction between the phenomenological and the scientific explanations of the physical world: a developmental study concerning observational astronomy. *Psychology* 11, 356–372 (in Greek).

Kyriakopoulou, N., and Vosniadou, S. (2008). The development of metaconceptual awareness in theory change. *Noisis* 4, 123–142 (in Greek).

Lombardi, D., and Sinatra, G. M. (2018). "Don't believe everything you think: reappraising judgements about conceptions," in *Converging Perspectives on Conceptual Change: Mapping an Emerging Paradigm in the Learning Sciences*, eds T. G. Amin and O. Levrini (New York, NY: Routledge), 237–244.

Mason, L. (2000). Role of anomalous data and epistemological beliefs in middle students' theory change on two controversial topics. *Eur. J. Psychol. Educ.* 15, 329–346. doi: 10.1007/bf03173183

Mason, L. (2002). "Developing epistemological thinking to foster conceptual change in different domains," in *Reconsidering Conceptual Change: Issues In Theory And Practice*, eds M. Limón and L. Mason (Dordrecht: Kluwer Academic Publishers), 301–335. doi: 10.1007/0-306-47637-1_16

Mason, L. (2003). "Personal epistemologies and intentional conceptual change," in *Intentional Conceptual Change*, eds G. M. Sinatra and P. R. Pintrich (Mahwah, NJ: Lawrence Erlbaum Associates), 204–241.

Meltzoff, A. N. (1995). Understanding the intentions of others: re-enactment of intended acts by 18-month-old children. *Dev. Psychol.* 31, 838–850. doi: 10.1037/0012-1649.31.5.838

Miller, S. A. (2012). *Theory of Mind: Beyond The Preschool Years*. London: Psychology Press.

Mitchell, P., and Lacohee, H. (1991). Children's early understanding of false belief. *Cognition* 39, 107–127. doi: 10.1016/0010-0277(91)90040-b

Moll, H., and Tomasello, M. (2007). Cooperation and human cognition: the Vygotskian intelligence hypothesis. *Philos. Trans. R. Soc. Lond. B* 362, 639–648. doi: 10.1098/rstb.2006.2000

Muris, P., Steerneman, P., Meesters, C., Merckelbach, H., Horselenberg, R., van den Hogen, T., et al. (1999). The TOM test: a new instrument for assessing theory of mind in normal children and children with pervasive developmental disorders. *J. Autism Dev. Disord.* 29, 67–80.

Osterhaus, C., Koerber, S., and Sodian, B. (2017). Scientific thinking in elementary school: children's social cognition and their epistemological understanding promote experimentation skills. *Dev. Psychol.* 53:450. doi: 10.1037/dev0000260

Perner, J. (1988). "Developing semantics for theories of mind: from propositional attitudes to mental representation," in *Developing Theories of Mind*, eds J. W. Astington, P. L. Harris, and D. R. Olson (Cambridge: Cambridge University Press), 141–172.

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Perner, J., and Wimmer, H. (1985). John thinks that mary thinks that. Attribution of second-order beliefs by 5 to 10 year old children. *J. Exper. Child Psychol.* 39, 437–471. doi: 10.1016/0022-0965(85)90051-7

Plummer, J. D. (2014). Spatial thinking as the dimension of progress in an astronomy learning progression. *Stud. Sci. Educ.* 50, 1–45. doi: 10.1080/03057267.2013.869039

Plummer, J. D., Bower, C. A., and Liben, L. S. (2016). The role of perspective taking in how children connect reference frames when explaining astronomical phenomena. *Intern. J. Sci. Educ.* 38, 345–365. doi: 10.1080/09500693.2016.1140921

Poulin-Dubois, D. (1999). "Infants' distinction between animate and inanimate objects: the origins of naive psychology," in *Early Social Cognition: Understanding Others In The First Months Of Life*, ed. P. Rochat (Hillsdale, NJ: Erlbaum), 257–280.

Pozo, J. I., Gomez, M. A., and Sanz, A. (1999). "When change does not mean replacement: different representations for different contexts," in *New Perspectives in Conceptual Change*, eds S. Wolfang, S. Vosniadou, and M. Carretero (Oxford: Elsevier-Pergamon), 161–173.

Qian, G., and Alvermann, D. (1995). Role of epistemological beliefs and learned helplessness in secondary school students' learning science concepts from text. *J. Educ. Psychol.* 87, 282–292. doi: 10.1037/0022-0663.87.2.282

Schwartz, R., and Brown, M. H. (2013). "Understanding photosynthesis and cellular respiration: encouraging a view of biological nested systems," in *Multiple Representations In Biological Education*, eds D. F. Treagust and C. Y. Tsui (Berlin: Springer Science & Business Media), 203–223. doi: 10.1007/978-94-007-4192-8_12

Seufert, T. (2003). Supporting coherence formation in learning from multiple representations. *Learn. Instr.* 13, 227–237. doi: 10.1016/s0959-4752(02)00022-1

Shtulman, A., and Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition* 124, 209–215. doi: 10.1016/j.cognition.2012.04.005

Smith, C. L., Maclin, D., Houghton, C., and Hennessey, M. G. (2000). Sixth graders' epistemologies of science: the impact of school science experiences on epistemological development. *Cogn. Instr.* 18, 349–422. doi: 10.1207/s1532690xci1803_3

Smith, C. L., and Wenk, L. (2006). Relations among three aspects of first-year college students' epistemologies of science. *J. Res. Sci. Teach.* 43, 747–785. doi: 10.1002/tea.20113

Sodian, B., and Kristen, S. (2016). "Theory of mind," in *Handbook of Epistemic Cognition*, eds J. A. Greene, W. A. Sandoval, and I. Bråten (Abingdon: Routledge), 80–97.

Sodian, B., Kristen-Antonow, S., and Koerber, S. (2016). Theory of mind predicts scientific reasoning. A longitudinal study from preschool to elementary school age. *Intern. J. Psychol.* 53, 360–360.

Sodian, B., Thoermer, C., Kircher, E., Grygier, P., and Günther, J. (2002). Vermittlung von wissenschaftsverständnis in der grundschule teaching understanding the nature of science in elementary school. *Zeitschrift fur Padagogik* 45, 192–206.

Spada, H. (1994). Conceptual change or multiple representations? *Learn. Instruct.* 4, 113–116. doi: 10.1016/0959-4752(94)90021-3

Taylor, M., Cartwright, B. S., and Bowden, T. (1991). Perspective taking and theory of mind: do children predict interpretive diversity as a function of differences in observers' knowledge? *Child Dev.* 62, 1334–1351.

Tomasello, M., and Moll, H. (2010). "The gap is social: human shared intentionality and culture," in *Mind the Gap*, eds P. M. Kappeler and J. B. Silk (New York, NY: Springer), 331–349. doi: 10.1007/978-3-642-02725-3_16

Treagust, D. F., Won, M., and McLure, F. (2017). "Multiple representations and students'conceptual change in science," in *Converging Perspectives On Conceptual Change: Mapping An Emerging Paradigm In The Learning Sciences*, eds T. G. Amin and O. Levrini (Abingdon: Routledge).

Vosniadou, S. (2019). The development of students' understanding of science. *Front. Educ.* 32. doi: 10.3389/feduc.2019.00032

Vosniadou, S., Archontidou, A., Kalogiannidou, A., and Ioannides, C. (1996). How greek children understand the shape of the earth: a study of conceptual change in childhood. *Psychol. Issues* 7, 30–51.

Vosniadou, S., and Brewer, W. F. (1992). Mental models of the earth: a study of conceptual change in childhood. *Cogn. Psychol.* 24, 535–585. doi: 10.1016/0010-0285(92)90018-W

Vosniadou, S., and Brewer, W. F. (1994). Mental models of the day/night cycle. *Cogn. Sci.* 18, 123–183.

Vosniadou, S., and Kyriakopoulou, N. (2006). "The problem of metaconceptual awareness in theory revision," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, eds R. Sun and N. Miyake (Greece: National Kapodistrian University of Athens University Campus), 2329–2334.

Vosniadou, S., and Skopeliti, I. (2014). Conceptual change from the framework theory side of the fence. *Sci. Educ.* 23, 1427–1445.

Vosniadou, S., Skopeliti, I., and Ikospentaki, K. (2004). Modes of knowing and ways of reasoning in elementary astronomy. *Cogn. Dev.* 19, 203–222.

Vosniadou, S., Pnevmatikos, D., Makris, N., Eikospentaki, K., Lepenioti, D., Chountala, A., et al. (2015). "Executive functions and conceptual change in science and mathematics learning," in *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, eds D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, and P. P. Maglio (Austin, TX: Cognitive Science Society), 2529–2534.

Vosniadou, S., Pnevmatikos, D., Makris, N., Lepenioti, D., Eikospentaki, K., Chountala, A., et al. (2018). The Recruitment of shifting and inhibition in on-line science and mathematics tasks. *Cogn. Sci.* 42, 1860–1886.

Wellman, H. M., Cross, D., and Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev.* 72, 655–684. doi: 10.1111/1467-8624.00304

White, S., Hill, E., Happé, F., and Frith, U. (2009). Revisiting the strange stories: revealing mentalizing impairments in autism. *Child Dev.* 80, 1097–1117. doi: 10.1111/j.1467-8624.2009.01319.x

Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13, 103–128. doi: 10.1016/0010-0277(83)90004-5

Wiser, M., and Amin, T. (2001). Is heat hot?" Inducing conceptual change by integrating everyday and scientific perspectives on thermal phenomena. *Learn. Instr.* 11, 331–355. doi: 10.1016/s0959-4752(00)00036-0

Wiser, M., and Smith, C. L. (2009). "Learning and teaching about matter in grades k–8: when should the atomic-molecular theory be introduced?," in *International Handbook Of Research On Conceptual Change*, ed. S. Vosniadou (Abingdon: Routledge), 233–267.

Won, M., Yoon, H., and Treagust, D. F. (2014). Students' learning strategies with multiple representations: explanations of the human breathing mechanism. *Sci. Educ.* 98, 840–866. doi: 10.1002/sce.21128

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Kyriakopoulou and Vosniadou. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 24 June 2020
doi: 10.3389/fpsyg.2020.01247

How Children's Cognitive Reflection Shapes Their Science Understanding

Andrew G. Young[1]* *and Andrew Shtulman*[2]

[1] Department of Psychology, Northeastern Illinois University, Chicago, IL, United States, [2] Department of Psychology, Occidental College, Los Angeles, CA, United States

Learning science requires contending with intuitions that are incompatible with scientific principles, such as the intuition that animals are alive but plants are not or the intuition that solids are composed of matter but gases are not. Here, we explore the tension between science and intuition in elementary school–aged children and whether that tension is moderated by children's tendency to reflect on their intuitions. Our participants were children between the ages of 5 and 12 years (n = 142). They were administered a statement-verification task, in which they judged statements about life and matter as true or false, as well as a children's Cognitive Reflection Test (CRT-D), in which they answered "brain teasers" designed to elicit an intuitive, yet inaccurate, response that could be corrected upon further reflection. Participants also received a tutorial on the scientific properties of life or matter, sandwiched between two blocks of the statement-verification task. We found that performance on the statement-verification task, which pitted scientific conceptions against intuitive conceptions (e.g., "cactuses are alive"), was predicted by performance on the CRT-D, independent of age. Children with higher levels of cognitive reflection verified scientific statements more accurately before the tutorial, and they made greater gains in accuracy following the tutorial. These results indicate that children experience conflict between scientific and intuitive conceptions of a domain in the earliest stages of acquiring scientific knowledge but can learn to resolve that conflict in favor of scientific conceptions, particularly if they are predisposed toward cognitive reflection.

Keywords: conceptual development, scientific reasoning, explanatory coexistence, intuitive theories, cognitive reflection

OPEN ACCESS

Edited by:
Ageliki Nicolopoulou, Lehigh University, United States

Reviewed by:
Deena Skolnick Weisberg, Villanova University, United States
Frank Keil, Yale University, United States

***Correspondence:**
Andrew G. Young
A-Young20@neiu.edu

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *02 February 2020*
Accepted: *13 May 2020*
Published: *24 June 2020*

Citation:
Young AG and Shtulman A (2020) How Children's Cognitive Reflection Shapes Their Science Understanding. Front. Psychol. 11:1247. doi: 10.3389/fpsyg.2020.01247

INTRODUCTION

Our first theories of natural phenomena – intuitive theories – are often incompatible with the scientific theories we learn later in life. They are developed by children from a combination of inputs, including innate biases, firsthand experience, and cultural teachings (Vosniadou, 1994; Carey, 2009; Shtulman, 2017), and they play the same inferential role as scientific theories, helping us explain past events, predict future events, and intervene on present events (Gopnik and Wellman, 2012). But they differ from scientific theories in that they carve up the world into entities and processes that do not align with the true causes of natural phenomena.

One well-studied example of intuitive theories is children's theories of life (Stavy and Wax, 1989; Hatano and Inagaki, 1994; Slaughter and Lyons, 2003). Life is a metabolic state – the consumption of energy to further an organism's survival and reproduction – but young children do not know

of the internal structures that make metabolism possible, so they interpret life as related to motion instead. Preschoolers classify moving but non-living entities, such as the sun and the clouds, as alive and classify living but non-moving objects, such as plants and trees, as not alive. These mistakes persist until children conceive of life as supported by the interrelated functions of internal organs, typically by age 10 years. Young children understand that organisms must eat and sleep in order to move and grow, but they lack the physioanatomical knowledge needed to conceive of them as bodily machines.

Another well-studied example is children's theories of matter (Carey, 1991; Nakhleh et al., 2005; Smith, 2007). Matter is anything composed of atoms, but most material substances betray no perceptible sign of their composition. Gases and vapors are all composed of atoms, but children can neither see them nor hold them, so they classify them as non-material. They also deny that such substances have weight or take up space. Children also make the converse mistake of classifying non-material entities that they can see or feel as matter, including echoes, shadows, and heat. This pattern persists until early adolescence, when children learn a particulate theory of matter in introductory physical science.

In the present study, we assess how children reason about life and matter in relation to their *cognitive reflection*, or their tendency to reflect on their own thinking. Learning about life and matter requires recognizing that one's intuitive understanding of a domain is incompatible with a scientific understanding, as well as the ability to suppress the former in favor of the latter. Children who are disposed to reflect on their own thinking may have an advantage at these tasks relative to those who are not. By studying how children's cognitive reflection relates to their understanding of counterintuitive scientific ideas, we shed new light on the domain-general resources that allow children to construct domain-specific theories of the natural world.

Explanatory Coexistence

Learning a scientific theory at odds with an intuitive theory requires conceptual change, or knowledge revision at the level of individual concepts. Conceptual change has traditionally been viewed as a process of restructuring and replacement (Carey, 1985; Nersessian, 1989; Chi, 1992; Vosniadou, 1994). Intuitive theories are restructured to accommodate counterintuitive scientific information and thus erased in the process, in the same way that remodeling a house erases the footprint of its original layout.

This view has been challenged by research revealing that intuitive theories continue to influence scientific reasoning throughout the life span, particularly when reasoners are cognitively burdened or cognitively impaired. In the domain of life, for instance, college undergraduates instructed to classify entities as "alive" or "not alive" under time pressure are prone to make the same mistakes that preschoolers make, classifying moving but non-living things as alive and living but non-moving things as not alive (Goldberg and Thompson-Schill, 2009). Alzheimer's disease patients make the same mistakes, even when given ample time to respond, and they explicitly define life in terms of motion rather than metabolic activity (Zaitchik and Solomon, 2008). Even elderly adults without Alzheimer's disease are inclined to conflate life with self-directed motion (Tardiff et al., 2017), indicating that childhood misconceptions persist across the life span and must be inhibited when reasoning about life as a metabolic process.

Early intuitions about matter also reemerge under cognitive load. Adults instructed to decide whether something is material or non-material as quickly as possible will mistakenly classify gases and heft-less objects, such as dust and snowflakes, as non-material and mistakenly classify perceptible forms of energy, such as rainbows and lightning, as material (Shtulman and Legare, 2020). Adults also make systematic mistakes in deciding whether an object will sink or float. When shown two balls of equal size, one made of wood and one made of lead, they judge that the wood ball is more likely to float than the lead one. But shown a large ball of wood and a small ball of lead, they take reliably longer to make the same judgment (Potvin et al., 2015; Potvin and Cyr, 2017).

Research over the past decade has revealed that this pattern is widespread (Shtulman and Lombrozo, 2016). Adults verify counterintuitive scientific ideas more slowly and less accurately than closely matched intuitive ones in several domains, including astronomy, genetics, mechanics, thermodynamics, and evolution (Shtulman and Valcarcel, 2012; Shtulman and Harrington, 2016). And these effects have been observed in several populations, including high schoolers (Babai et al., 2010), undergraduate science majors (Foisy et al., 2015), high school science teachers (Potvin and Cyr, 2017), and elderly adults (Barlev et al., 2018). Even professional physicists (Kelemen et al., 2013) and professional biologists (Goldberg and Thompson-Schill, 2009) exhibit cognitive conflict when reasoning about counterintuitive scientific ideas. Such conflict indicates that early intuitions about natural phenomena survive the acquisition of contradictory scientific knowledge.

Cognitive Reflection

The current study investigates whether learning counterintuitive scientific ideas is shaped by cognitive reflection or the disposition to reflect on, and override, our first intuition. This disposition is most commonly measured by the Cognitive Reflection Test (CRT; Frederick, 2005), a three-item test designed for adults. Consider this item: "In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?" The correct answer is 47, given the patch must have covered half the lake a day prior to covering the entire lake, but the question is designed to elicit an intuitive response of 24, or half of 48.

Adults who perform well on the CRT demonstrate superior performance on many other reasoning tasks, including those measuring logical reasoning, probabilistic reasoning, argumentation, and temporal discounting (Frederick, 2005; Toplak et al., 2011). The CRT is a stronger predictor of performance on such tasks than either general intelligence or executive functioning (Toplak et al., 2011). The CRT also correlates with causal reasoning (Don et al., 2016), moral reasoning (Royzman et al., 2014), endorsement of scientific claims (Gervais, 2015), rejection of supernatural

claims (Pennycook et al., 2012), rejection of stereotypes (Hammond and Cimpian, 2017), and detection of fake news (Pennycook and Rand, 2018).

Cognitive reflection has been studied extensively in adults, but little is known about its development. To address this gap, we created a nine-item CRT for school-aged children, the CRT–Developmental Version (CRT-D). Each item elicits an intuitive, but incorrect, lure response that school-aged children should be able to correct upon further reflection. In a preliminary study (Young A. G. et al., 2018), we found that adults' performance on the CRT-D was strongly correlated with their performance on the original CRT, as well as their performance on various heuristics-and-biases tasks, and children's performance on the CRT-D was strongly correlated with child-friendly versions of the same tasks, even when controlling for age.

Research Objectives

Prior research indicates that cognitive reflection supports science understanding in adults (Shtulman and McCallum, 2014). College students' CRT scores predict their understanding of astronomy, evolution, geology, mechanics, perception, and thermodynamics more strongly than their prior STEM coursework, their statistical reasoning ability, or their understanding of the nature of science. Here, we explore whether CRT scores predict science understanding in children, who are in the earliest stages of learning science and who have less experience reflecting on their own cognition.

We also explore whether cognitive reflection predicts children's ability to learn new scientific information, by providing them with tutorials on life and matter. In previous research (Young A. et al., 2018), we found that such tutorials promote adults' ability to verify counterintuitive scientific ideas. Adults are reliably slower and less accurate at verifying counterintuitive statements such as "dust has weight" and "yeast needs nutrients" relative to intuitive statements involving the same predicates, such as "bricks have weight" and "goats need nutrients." Providing adults with tutorials on the scientific properties of life and matter helped them close the gap in accuracy between the two types of statements but not the gap in latency. In other words, the tutorials did not assuage the immediate conflict elicited by counterintuitive statements (as indexed by response times), but they did help participants favor scientific responses over intuitive ones.

In the present study, we extended this line of research to elementary schoolers. Our study followed the same protocol as Young A. et al. (2018), which included a pretest, a tutorial, and a posttest. It expanded on that protocol by including a domain-general measure of cognitive ability, the CRT-D. We expected children to show signs of conflict between science and intuition, given that the children in our age range had begun to learn about life and matter in school, and we expected children to verify counterintuitive statements more accurately following instruction. It was an open question, though, whether children's performance on the statement-verification task would correlate with their performance on the CRT-D or whether improvements in performance, from pretest to posttest, would correlate as well.

METHODS

Participants

Our participants were 142 children in kindergarten through 6th grade. Their mean age was 8 years and 5 months, and they were approximately balanced for gender (82 female, 62 male). Children were recruited from public playgrounds and a children's museum, and they completed the study onsite.

Materials

Cognitive Reflection Test – Developmental Version

Children answered the nine cognitive reflection questions in **Table 1** (from Young A. G. et al., 2018). We used the number of correct responses as children's score, with higher scores indicating greater cognitive reflection (mean = 2.8, *SD* = 1.9, range = 0–8). Reliability for the measure was acceptable (McDonald's ω total = 0.74). While some children may have lacked the knowledge or cultural background required to answer certain items correctly, we took a conservative approach and retained all CRT-D items because they matched the response structure of the original CRT, namely, they elicited more intuitive responses than other incorrect responses. Ongoing research aims to verify that CRT-D items are functioning as intended across diverse samples.

Statement-Verification Task

We measured children's understanding of counterintuitive scientific ideas with a statement-verification task. Children were asked to judge four types of statements as true or false. Some statements were true from both a scientific perspective and an intuitive perspective ("tigers need nutrients"); some were false from both perspectives ("forks need nutrients"); some were true from a scientific perspective but false from an intuitive

TABLE 1 | Items on the Cognitive Reflection Test–Developmental Version (CRT-D), along with their correct answer and the intuitive answer they were designed to prime.

Item	Correct	Intuitive
If you're running a race and you pass the person in second place, what place are you in?	Second	First
Emily's father has three daughters. The first two are named Monday and Tuesday. What is the third daughter's name?	Emily	Wednesday
A farmer has 5 sheep, all but 3 run away. How many are left?	Three	Two
If there are 3 apples and you take away 2, how many do you have?	Two	One
What do cows drink?	Water	Milk
What weighs more, a pound of rocks or a pound of feathers?	Same weight	Rocks
What hatches from a butterfly egg?	Caterpillar	Butterfly
Who makes Christmas presents at the North Pole?	Elves	Santa
Anna is playing foursquare with her three friends: Eeny, Meeny, and Miny. Who is the fourth player?	Anna	Mo

perspective ("bacteria need nutrients"), and some were false from a scientific perspective but true from an intuitive perspective ("fire needs nutrients"). The first two types of statements will be referred to as intuitive, and the latter two types as counterintuitive.

For each domain, statements were generated by pairing three predicates with 32 entities. In the domain of life, the predicates were "reproduces," "needs nutrients," and "grows and develops." These predicates apply to all living things, but we predicted children would be more inclined to apply them to entities that exhibit self-directed motion. In the domain of matter, the predicates were "has weight," "takes up space," and "is made of atoms." These predicates apply to all material things, but we predicted children would be more inclined to apply them to entities that can be seen or felt. Because children might not know the meanings of certain predicates, we defined each on first introduction. "Reproduce," for instance, was defined as "things that can make more things like themselves," and "made of atoms" was defined as "things that are made of up of tiny pieces."

Predicates were paired with the four types of entities shown in **Table 2**. In the domain of life, those entities were animals (deemed alive by both science and intuition), inanimate artifacts and inanimate natural kinds (deemed alive by neither science nor intuition), plants and microorganisms (deemed alive by science but not intuition), and animate natural kinds (deemed alive by intuition but not science). In the domain of matter, those entities were physical objects (deemed material by both science and intuition), abstract ideas (deemed material by neither science nor intuition), gases and other substances lacking bulk or heft (deemed material by science but not intuition), and the visible or tangible components of energy (deemed material by intuition but not science). These pairings created the four types of statements described above: statements deemed true by both science and intuition ("bricks take up space"), statements deemed false by both science and intuition ("dreams take up space"), statements deemed true by science but not intuition ("air takes up space"), and statements deemed true by intuition but not science ("rainbows take up space").

Children completed the task on an iPad, responding via touch screen. Fifty children opted into a version of the task that played audio recordings of the statements, obviating the need to read the statements. Audio recordings of each statement were generated using Apple's macOS text-to-speech engine. Children who listened to the audio-recorded stimuli received only four of the six predicates, due to the additional time required to play the recordings.

TABLE 2 | Sample items used in the biological statements (top) and physical statements (bottom) on the statement-verification task, organized by their classification according to science and intuition.

	Intuitive classification	
Scientific classification	**Living**	**Non-living**
Living	Rabbits	Mushrooms
	Turtles	Grass
	Snails	Bacteria
Non-living	Sun	Hammers
	Wind	Caves
	Fire	Shells
	Intuitive classification	
Scientific classification	**Material**	**Non-material**
Material	Bricks	Smoke
	Ice	Snowflakes
	Logs	Air
Non-material	Rainbows	Dreams
	Shadows	Songs
	Heat	Numbers

Tutorials

Children completed a tutorial on life or matter midway through the experiment. Each tutorial began with definitions of key characteristics of the domain, followed by a brief video that illustrated those characteristics with examples. The tutorials then addressed common misconceptions about the domain, followed by videos that illustrated why these misconceptions were false. The tutorial on life emphasized that all living things need energy and nutrients, grow and develop, react to stimuli in their environment, and reproduce. It addressed the misconception that life is synonymous with self-directed motion with examples of entities that do not move but are alive (e.g., moss) and entities that move but are not alive (e.g., comets). The tutorial on matter emphasized that all matter occupies space, has weight, is made of atoms, and can undergo phase transitions. It addressed the misconception that matter is synonymous with perceptibility with examples of entities that cannot be perceived but are material (e.g., gases) and entities that can be perceived but are not material (e.g., lightning). Tutorials took approximately 7 min to complete.

Procedure

Children completed the CRT-D, then verified 48 statements about life and 48 statements about matter (pretest), then completed a tutorial on life or matter, and finally verified 48 additional statements from each domain (posttest). Children were randomized to tutorial condition in equal proportions.

Children completed the pretest and posttest in blocks. They saw a screen introducing a particular predicate ("Does it grow and develop?"), followed by 16 statements with that predicate ("Seaweed grows and develops"). The statements were randomly ordered within a block, and the blocks were randomly ordered within the testing phase, meaning that biological and physical predicates were intermixed.

Children saw the same predicates at pretest and posttest, but those predicates were paired with 16 new items; that is, the 48 statements children saw at pretest (3 predicates × 16 items) were different from the 48 statements they saw at posttest. Items were randomly assigned to one of two item sets for counterbalancing, so that the 48-statement pretest for some children constituted the 48-statement posttest for other children and *vice versa*. Because of experimenter error, item sets were imperfectly balanced. One

set was presented at pretest for 73% of children, whereas the other was presented at pretest for 27% of children. Preliminary analyses found item set was not a significant predictor of speed or accuracy, by itself or in interaction with statement type, in either domain at either test period, indicating that children's performance was consistent across item sets.

RESULTS

The statement-verification task yielded two outcome measures: response accuracy and response latency. We analyzed each outcome with a linear mixed model (LMM), with statement type (intuitive or counterintuitive), test (pretest or posttest), instruction (instructed or uninstructed), and their interactions as fixed effects and by-participant and by-predicate random effects. Models with maximal random effects structures had convergence issues, and thus we followed the procedure recommended by Bates et al. (2015) to guide removal of random effects that were not supported by the data. Inference for fixed effects was carried out via type 3 likelihood ratio test (LRT) model comparison.

Our primary analyses collapse across tutorial domain (life or matter) and focus on whether the statements were targeted by instruction or not. Response latencies were similar across domains (2.83 vs. 2.93 s), and effects of instruction were similar across domains. But children did verify biological statements more accurately than physical statements (85 vs. 74%), so we report domain-specific results when analyzing response accuracy.

As noted above, some children listened to the scientific statements, and some read them on their own. The latter responded more quickly (2.30 vs. 3.96 s) and more accurately (82 vs. 74%) and were also older (9;1 vs. 7;2) and higher in cognitive reflection (3.2 vs. 2.1). However, a parallel set of models that adjusted for presentation format (read or listen) and its interactions with other predictors yielded similar findings to the reported results.

Response Accuracy

As seen in **Figure 1**, there was an effect of statement type, such that children verified intuitive statements more accurately than counterintuitive statements, LRT $\chi^2(1) = 10.46$, $p < 0.001$. Overall, accuracy for intuitive statements was 18.5% greater than accuracy for counterintuitive statements, 95% confidence interval (CI) [12.1, 24.9]. Additionally, there was a three-way interaction between statement type, test period, and instruction, LRT $\chi^2(1) = 11.15$, $p < 0.001$. In the instructed domain, children's posttest accuracy for counterintuitive statements was 9.4% greater than their pretest performance, 95% CI [7.1, 11.6]. This learning was observed in both the life domain, 95% CI [1.7, 8.1], and matter domain, 95% CI [10.6, 16.8]. This effect was limited to counterintuitive statements in the instructed domain; intuitive statements were verified with similar accuracy at pretest and posttest in both domains, as were counterintuitive statements in the uninstructed domain. Instruction was thus effective at improving children's accuracy at verifying counterintuitive scientific ideas within the targeted domain.

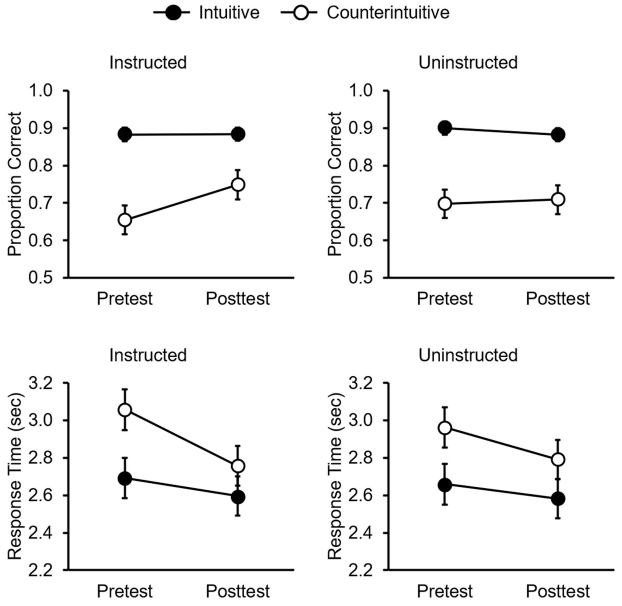


FIGURE 1 | Estimated proportion of correct verifications **(top)** and response latency **(bottom)** by statement type, test, and instruction. Error bars represent standard errors.

Response Latency

Following prior research, we analyzed response latencies for correctly verified statements only. Response latency thus indicates whether arriving at a correct response entailed more cognitive conflict for some statements relative to others. Before doing so, we first removed latencies shorter than 250 ms, as they were too quick to have been deliberate. We then calculated the mean response latency across participants and statements (mean = 2,968 ms) and removed latencies more than 2 SDs above the mean (>9,071 ms).

As seen in **Figure 1**, there was an effect of statement type, such that children correctly verified counterintuitive statements more slowly than intuitive ones, LRT $\chi^2(1) = 102.61$, $p < 0.001$. Response latencies for counterintuitive statements were 260 ms slower than response latencies for intuitive statements, 95% CI [210, 309]. Additionally, there was an interaction between test and statement type, LRT $\chi^2(1) = 8.79$, $p = 0.003$. Children correctly verified counterintuitive statements 198 ms faster at posttest than pretest, 95% CI [43, 352], but response latencies for intuitive statements were similar at pretest and posttest. We suspect this effect was due to increased familiarity with the task and greater initial latencies at pretest, as it did not vary by instruction [three-way interaction: LRT $\chi^2(1) = 1.18$, $p = 0.278$].

Cognitive Reflection

Children's CRT-D performance was moderately to strongly correlated with response accuracy at both test periods for both types of statements in both domains (r's = 0.27–0.53, p's < 0.002). These correlations indicate that children with higher CRT-D scores performed more accurately on the statement-verification

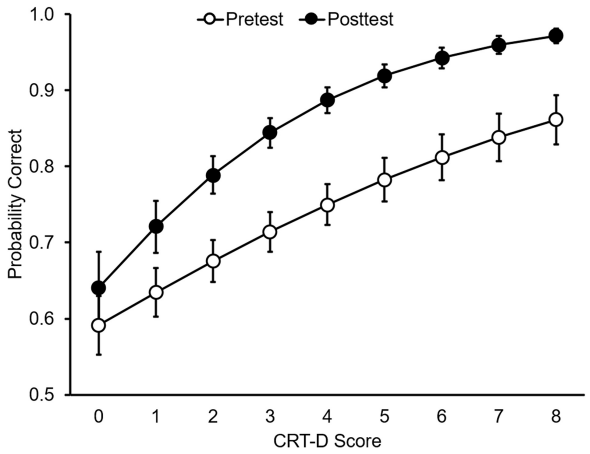


FIGURE 2 | Estimated probability of correct verifications by CRT-D score and test. Error bars represent standard errors.

task across the board, but did they also learn more from instruction?

We estimated a binomial generalized LMM on children's correct responses for counterintuitive statements in the instructed domain with test (pretest or posttest), CRT-D score, and their interaction as fixed effects and by-participant and by-item random effects. This analysis revealed an interaction between test and CRT-D score, LRT $\chi^2(1) = 12.19$, $p < 0.001$. Children with higher CRT-D scores showed larger gains in accuracy from pretest to posttest, logit $b_{Test \times CRT.D} = 0.19$, 95% CI [0.08, 0.29], as shown in **Figure 2**. This result was observed in both the life domain, 95% CI [0.00, 0.32], and matter domain, 95% CI [0.08, 0.34]. Critically, the interaction between test and CRT-D score remained significant in an additional model that included fixed effects for age and an age-by-test interaction, LRT $\chi^2(1) = 4.30$, $p = 0.038$.

Children's CRT-D performance yielded moderate to strong negative correlations with response latencies at both test periods for both types of statements in both domains (r's = −0.39 to −0.48, p's < 0.001). We explored the potential effects of CRT-D performance by estimating an LMM with statement type (intuitive or counterintuitive), test (pretest or posttest), instruction (instructed or uninstructed), CRT-D, and their interactions as fixed effects and by-participant and by-predicate random effects. There was an overall effect of CRT-D performance, LRT $\chi^2(1) = 24.14$, $p < 0.001$, such that children with greater cognitive reflection had shorter latencies, but no interactions involving CRT-D were observed.

DISCUSSION

Do elementary schoolers exhibit cognitive conflict when reasoning about counterintuitive scientific ideas? And does their tendency to reflect on their own cognition moderate this conflict? Our findings support both possibilities. Children were slower and less accurate at verifying scientific statements that conflict with their intuitive theories of life or matter compared to closely matched statements that accord with those theories. Instructing children on the scientific properties of life or matter increased their accuracy for counterintuitive statements in the instructed domain but did not increase their speed (relative to intuitive statements). These findings indicate that children experience conflict between scientific ideas and intuitive ideas, despite limited exposure to science, but this conflict can be resolved in favor of scientific ideas with targeted instruction.

Children's accuracy at verifying domain-specific scientific statements was predicted by a domain-general disposition: cognitive reflection. Cognitively reflective children were more accurate at verifying scientific statements at both pretest and posttest. They also learned more from instruction, exhibiting larger gains in accuracy from pretest to posttest than children with lower CRT-D scores.

Our findings parallel those of Young A. et al. (2018), who administered the same task to adults. The adults were faster and more accurate than children in the present study, but both groups verified counterintuitive statements more slowly and less accurately than closely matched intuitive statements. The effect of instruction was also similar across age groups, increasing participants' accuracy at verifying counterintuitive statements but not their speed. Cognitive conflict between science and intuition thus appears to take the same form across development.

Our findings also parallel those documented by Vosniadou et al. (2018), who assessed tensions between children's intuitive and scientific reasoning in a different task. These researchers asked third- and fifth-graders to sort physical and biological items into one of two categories: a category that emphasized the item's intuitive features or a category that emphasized its scientific features. For instance, participants decided whether water should be grouped with other liquids (coke, lemonade, milk) or with other forms of H_2O (ice, vapor, snow). Children of all ages preferred intuitive categories over scientific categories, and they took longer to make their judgments when they opted for the scientific category instead.

Vosniadou et al. also measured children's executive function skills – set-shifting ability and inhibitory control – and found that children with higher executive function were more likely to categorize the target items by their scientific properties and were faster to do so (see also Bascandziev et al., 2018; Tardiff et al., 2020). These findings echo our finding that children with higher cognitive reflection were more accurate at verifying counterintuitive scientific ideas, and they raise questions about the relation between cognitive reflection and executive function. Cognitive reflection draws on similar skills, as children must inhibit a gut response (inhibition) in order to shift to another response (set shifting) while holding the question in mind (working memory). But succeeding on the CRT-D also requires recognizing that such activities are necessary, as well as the ability to coordinate them on one's own. The "stop and think" aspect of cognitive reflection may transcend the individual components of executive function. Studies of rational thought have found that cognitive reflection predicts performance on

heuristics-and-biases tasks independent of executive function (Toplak et al., 2011; Young and Shtulman, 2020), but future research is needed to determine whether the same is true of science understanding and, if so, which aspects of science understanding are uniquely predicted by cognitive reflection.

The current findings suggest that cognitive reflection may be a prerequisite for changing certain cognitive representations, but it remains unclear as to why. Cognitively reflective individuals may be better at identifying gaps in their understanding, or they may be better at filling those gaps with new information. They may be more receptive to instruction, or they may be better at monitoring and resolving response conflicts. We suspect that cognitive reflection is valuable because it fosters metaconceptual awareness. All children reason *with* their concepts, but they might not reason *about* their concepts, and this latter ability may be required for changing them. Pedagogically, our findings imply that instructors could use the CRT-D as a diagnostic for determining who is likely to profit from instruction and who is not. Children with low CRT-D scores may benefit from more instruction, or different instruction, than their peers.

In conclusion, we have shown that conflict between science and intuition emerges early in the acquisition of scientific knowledge. Children in the earliest stages of science education verify scientific ideas that conflict with their intuitive theories more slowly and less accurately than those that accord with them. Although this conflict may be inevitable, children can learn to privilege scientific ideas over intuitive ones with instruction that challenges intuitive theories and a disposition toward questioning intuitive responses.

DATA AVAILABILITY STATEMENT

The datasets for this study are available at https://osf.io/rxw27/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Occidental College Institutional Review Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

Both authors contributed to all aspects of the data collection, the data analysis, and the write up.

ACKNOWLEDGMENTS

We would like to thank Isabel Geddes and Claire Weider for assistance with the data collection and data coding. We would also like to thank the James S. McDonnell Foundation for supporting this research with an Understanding Human Cognition Scholar Award to AS.

REFERENCES

Babai, R., Sekal, R., and Stavy, R. (2010). Persistence of the intuitive conception of living things in adolescence. *J. Sci. Educ. Technol.* 19, 20–26. doi: 10.1007/s10956-009-9174-2

Barlev, M., Mermelstein, S., and German, T. C. (2018). Representational coexistence in the God concept: core knowledge intuitions of God as a person are not revised by Christian theology despite lifelong experience. *Psychonom. Bull. Rev.* 25, 2330–2338. doi: 10.3758/s13423-017-1421-6

Bascandziev, I., Tardiff, N., Zaitchik, D., and Carey, S. (2018). The role of domain-general cognitive resources in children's construction of a vitalist theory of biology. *Cogn. Psychol.* 104, 1–28. doi: 10.1016/j.cogpsych.2018.03.002

Bates, D., Kliegl, R., Vasishth, S., and Baayen, R. H. (2015). Parsimonious mixed models. *arXiv [Preprint]* Available at: https://arxiv.org/pdf/1506.04967.pdf (accessed January 1, 2020).

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Carey, S. (1991). "Knowledge acquisition: enrichment or conceptual change," in *The Epigenesis of Mind: Essays in Biology and Cognition*, eds S. Carey, and R. Gelman, (Hillsdale, NJ: Lawrence Erlbaum), 257–291.

Carey, S. (2009). *The Origin of Concepts*. New York, NY: Oxford University Press.

Chi, M. (1992). "Conceptual change within and across ontological categories: examples from learning and discovery in science," in *Cognitive Models of Science*, ed. R. Giere, (Minneapolis, MN: University of Minnesota Press), 129–186.

Don, H. J., Goldwater, M. B., Otto, A. R., and Livesey, E. J. (2016). Rule abstraction, model-based choice, and cognitive reflection. *Psychonom. Bull. Rev.* 23, 1615–1623. doi: 10.3758/s13423-016-1012-y

Foisy, L. M. B., Potvin, P., Riopel, M., and Masson, S. (2015). Is inhibition involved in overcoming a common physics misconception in mechanics? *Trends Neurosci. Educ.* 4, 26–36. doi: 10.1016/j.tine.2015.03.001

Frederick, S. (2005). Cognitive reflection and decision making. *J. Econ. Perspect.* 19, 25–42. doi: 10.1257/089533005775196732

Gervais, W. M. (2015). Override the controversy: analytic thinking predicts endorsement of evolution. *Cognition* 142, 312–321. doi: 10.1016/j.cognition.2015.05.011

Goldberg, R. F., and Thompson-Schill, S. L. (2009). Developmental "roots" in mature biological knowledge. *Psychol. Sci.* 20, 480–487. doi: 10.1111/j.1467-9280.2009.02320.x

Gopnik, A., and Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychol. Bull.* 138, 1085–1108. doi: 10.1037/a0028044

Hammond, M. D., and Cimpian, A. (2017). Investigating the cognitive structure of stereotypes: generic beliefs about groups predict social judgments better than statistical beliefs. *J. Exp. Psychol.* 146, 607–614. doi: 10.1037/xge0000297

Hatano, G., and Inagaki, K. (1994). Young children's naive theory of biology. *Cognition* 50, 171–188. doi: 10.1016/0010-0277(94)90027-2

Kelemen, D., Rottman, J., and Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: purpose-based reasoning as a cognitive default. *J. Exp. Psychol.* 142, 1074–1083. doi: 10.1037/a0030399

Nakhleh, M. B., Samarapungavan, A., and Saglam, Y. (2005). Middle school students' beliefs about matter. *J. Res. Sci. Teach.* 42, 581–612. doi: 10.1002/tea.20065

Nersessian, N. J. (1989). Conceptual change in science and in science education. *Synthese* 80, 163–183.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., and Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition* 123, 335–346. doi: 10.1016/j.cognition.2012.03.003

Pennycook, G., and Rand, D. G. (2018). Lazy, not biased: susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188, 39–50. doi: 10.1016/j.cognition.2018.06.011

Potvin, P., and Cyr, G. (2017). Toward a durable prevalence of scientific conceptions: tracking the effects of two interfering misconceptions about buoyancy from preschoolers to science teachers. *J. Res. Sci. Teach.* 54, 1121–1142. doi: 10.1002/tea.21396

Potvin, P., Masson, S., Lafortune, S., and Cyr, G. (2015). Persistence of the intuitive conception that heavier objects sink more: a reaction time study with different levels of interference. *Int. J. Sci. Math. Educ.* 13, 21–43. doi: 10.1007/s10763-014-9520-6

Royzman, E. B., Landy, J. F., and Leeman, R. F. (2014). Are thoughtful people more utilitarian? CRT as a unique predictor of moral minimalism in the dilemmatic context. *Cogn. Sci.* 39, 325–352. doi: 10.1111/cogs.12136

Shtulman, A. (2017). *Scienceblind: Why Our Intuitive Theories About the World Are so Often Wrong*. New York, NY: Basic Books.

Shtulman, A., and Harrington, K. (2016). Tensions between science and intuition across the lifespan. *Top. Cogn. Sci.* 8, 118–137. doi: 10.1111/tops.12174

Shtulman, A., and Legare, C. H. (2020). Competing explanations of competing explanations: accounting for conflict between scientific and folk explanations. *Top. Cogn. Sci.* doi: 10.1111/tops.12483

Shtulman, A., and Lombrozo, T. (2016). "Bundles of contradiction: a coexistence view of conceptual change," in *Core Knowledge and Conceptual Change*, eds D. Barner, and A. Baron, (Oxford: Oxford University Press), 49–67.

Shtulman, A., and McCallum, K. (2014). "Cognitive reflection predicts science understanding," in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, eds P. Bello, M. Guarini, M. McShane, and B. Scassellati, (Austin, TX: Cognitive Science Society), 2937–2942.

Shtulman, A., and Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition* 124, 209–215. doi: 10.1016/j.cognition.2012.04.005

Slaughter, V., and Lyons, M. (2003). Learning about life and death in early childhood. *Cogn. Psychol.* 46, 1–30. doi: 10.1016/s0010-0285(02)00504-2

Smith, C. L. (2007). Bootstrapping processes in the development of students' commonsense matter theories: using analogical mappings, thought experiments, and learning to measure to promote conceptual restructuring. *Cogn. Instr.* 25, 337–398. doi: 10.1080/07370000701632363

Stavy, R., and Wax, N. (1989). Children's conceptions of plants as living things. *Hum. Dev.* 32, 88–94. doi: 10.1159/000276367

Tardiff, N., Bascandziev, I., Carey, S., and Zaitchik, D. (2020). Specifying the domain-general resources that contribute to conceptual construction: evidence from the child's acquisition of vitalist biology. *Cognition* 195:104090. doi: 10.1016/j.cognition.2019.104090

Tardiff, N., Bascandziev, I., Sandor, K., Carey, S., and Zaitchik, D. (2017). Some consequences of normal aging for generating conceptual explanations: a case study of vitalist biology. *Cogn. Psychol.* 95, 145–163. doi: 10.1016/j.cogpsych.2017.04.004

Toplak, M. E., West, R. F., and Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Mem. Cogn.* 39, 1275–1289. doi: 10.3758/s13421-011-0104-1

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learn. Instr.* 4, 45–69. doi: 10.1016/0959-4752(94)90018-3

Vosniadou, S., Pnevmatikos, D., Makris, N., Lepenioti, D., Eikospentaki, K., Chountala, A., et al. (2018). The recruitment of shifting and inhibition in online science and mathematics tasks. *Cogn. Sci.* 42, 1860–1886. doi: 10.1111/cogs.12624

Young, A., Laca, J., Dieffenbach, G., Hossain, E., Mann, D., and Shtulman, A. (2018). "Can science beat out intuition? Increasing the accessibility of counterintuitive scientific ideas," in *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, Madison, 1238–1243.

Young, A. G., Powers, A., Pilgrim, L., and Shtulman, A. (2018). "Developing a cognitive reflection test for school-age children," in *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, eds T. T. Rogers, M. Rau, X. Zhu, and C. W. Kalish, (Austin, TX: Cognitive Science Society), 1232–1237.

Young, A. G., and Shtulman, A. (2020). Children's cognitive reflection predicts conceptual understanding in science and mathematics. *Psychol. Sci.* (in press).

Zaitchik, D., and Solomon, G. E. A. (2008). Animist thinking in the elderly and in patients with Alzheimer's disease. *Cognitive Neuropsychology* 25, 27–37. doi: 10.1080/02643290801904059

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Young and Shtulman. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 26 June 2020
doi: 10.3389/fpsyg.2020.01475

Young Schoolchildren's Epistemic Development: A Longitudinal Qualitative Study

Michael Weinstock, Vardit Israel, Hadas Fisher Cohen, Iris Tabak and Yifat Harari*

Department of Education, Ben-Gurion University of the Negev, Beersheba, Israel

How children seek knowledge and evaluate claims may depend on their understanding of the source of knowledge. What shifts in their understandings about why scientists might disagree and how claims about the state of the world are justified? Until about the age of 41/2, knowledge is seen as self-evident. Children believe that knowledge of reality comes directly through our senses and what others tell us. They appeal to these external sources in order to know. The attainment of Theory of Mind (ToM) at this age is commonly seen as the significant shift in development in understanding disagreements in knowledge claims. Children attaining ToM understand that someone exposed to incorrect or incomplete information might have false beliefs. Disagreement, then, is still attributed to objective sources of knowledge. The current study examines the later developing Interpretive Theory of Mind (iToM) as the basis for children's understanding of how people with access to the same information might disagree and what this means for how to provide justification for a knowledge claim. Fourteen 2nd graders with the most iToM responses to four tasks and 14 with the fewest iToM responses were selected from a larger sample of 91. In analyses of interviews about a story in which two experts make different claims about a scientific phenomenon, those in the high iToM group noted subjective perspective and processes as the source of disagreement and suggested the need for investigation as the means to knowing. In contrast, those in the low iToM group mostly could not explain the source of disagreement and held that knowledge is acquired from external sources. A comparison of the interviews regarding the science story 2 years later allows for a qualitative description of the development. Those in the low iToM group showed more general recognition of subjective and constructive processes in knowing whereas those in the high iToM group identified interpretive processes and the relativity of perspectives with implications for how observations were conducted and interpreted. Only those in the high iToM group referred to the importance of evidence as a basis for knowledge claims at either point in the study.

Keywords: epistemic understanding, epistemic development, interpretive theory of mind, knowledge justification, source of knowledge, discrepant claims

OPEN ACCESS

Edited by:
Christopher Osterhaus, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Emanuela Yeung, University of Copenhagen, Denmark
Beate Sodian, Ludwig Maximilian University of Munich, Germany

****Correspondence:***
Michael Weinstock
micwein@bgu.ac.il

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *31 January 2020*
Accepted: *02 June 2020*
Published: *26 June 2020*

Citation:
Weinstock M, Israel V, Cohen HF, Tabak I and Harari Y (2020) Young Schoolchildren's Epistemic Development: A Longitudinal Qualitative Study. Front. Psychol. 11:1475. doi: 10.3389/fpsyg.2020.01475

INTRODUCTION

According to Kuhn (2001) and Kuhn et al. (2008), the separation and then coordination of theory and evidence are the essential bases for constructing claims in science and other knowledge domains. Epistemic understandings regarding the nature of theory and evidence as well as understandings of the standards of knowledge justification and the sources of knowledge underlie the competent performance of these tasks (Kuhn, 1991). Research has found that epistemic understandings are related to knowledge construction, justification, and evaluation among older school-age children (Mason et al., 2008; Iordanou, 2010; Barzilai and Zohar, 2012), adolescents (Mason and Scirica, 2006; Weinstock et al., 2006), and adults (Nussbaum et al., 2008; Kuhn et al., 2010; Weinstock, 2016).

Although younger school-age children's epistemic understandings have been inferred from their behavior and responses to problems (e.g., Sodian et al., 1991; Pillow, 2002; Sandoval and Çam, 2011; Huang et al., 2019), with few exceptions (e.g., Ryu and Sandoval, 2012; Osterhaus et al., 2017) there has been little examination of their explicit epistemic understandings of theory, evidence, justification, and source of knowledge and of how they understand the epistemic characteristics of others' thinking. Similarly, in research referring specifically to people's understandings of the nature of science (NOS) there have been few studies on young school children and they tend to focus more on science reasoning, strategies, and use of theories and evidence rather than looking at their explicit understandings (Metz, 2004; Osterhaus et al., 2017). Koerber et al. (2015) did find that a significant percentage, although not a majority, of 9-year-olds did have understanding of at least one aspect of NOS. This shows that there is increasing understanding with age. But they also stressed that they could not identify a particular trajectory in the development of understanding of NOS. The current research is a step in trying to identify the trajectory of younger school-age children's explicit epistemic understandings.

Although children's verbal abilities might limit their meta-epistemic expression (Kloo et al., 2017; Huang et al., 2019), much research with adolescents and adults has shown that confronting discrepant knowledge makes epistemic perspective more salient (Perry, 1970; King and Kitchener, 1994; Kuhn and Weinstock, 2002; Barzilai and Weinstock, 2015). This seems also true of children as many theory-of-mind related tasks use competing claims or perspectives (Wellman et al., 2011; Osterhaus et al., 2016). The current research employs this discrepant-claims method to encourage children to express their epistemic assumptions. In short, we are interested in how schoolchildren understand and explain the source of disagreement between science experts and the means of resolving disagreement or uncertainty through justification. Although theory of mind (ToM) research is replete with examples of how children understand that different minds might know different things or how someone might lack knowledge (Wellman et al., 2011), we are interested in how theories of mind begin to be transformed into more generalized theories of knowledge, from understanding that someone with different information might have a false belief to understanding how it is that knowledge is not self-evident and that people might develop different claims based on the same information.

Much of the research on adolescents' and adults' epistemic thinking, such as the studies cited above, refer to a common description of epistemic development (Kuhn and Weinstock, 2002) which proposes a course of development through several epistemic perspectives. The developmental task that produces such development is to reconcile the objective and subjective aspects of knowledge. People progress from an "absolutist" understanding that knowledge is objective, certain, and in no need of justification, and has reality or authority as its source. With enough experience of uncertainty in knowledge and disagreement, even among experts, about what is known, people develop a "multiplist" understanding that there are multiple possible perspectives on truth. Knowledge is seen as subjective and basically uncertain because people do not agree with one another. Although there are means to explain one's position, justification will not be fruitful because the source of claims are individual opinions, and thus, competing claims cannot be adjudicated. The succeeding "evaluativist" epistemic perspective is a shift from the radical, opinion-oriented relativism of the multiplist perspective while still understanding that knowledge may be relative to the knower and there may be different perspectives on truth. In this perspective, it is understood that judgments must and can be supported through processes of justification. It is understood that reality does not force knowledge on us, but that information must be interpreted in order to construct knowledge. Knowledge is not certain, but through the use of sound methods with clear standards of knowledge justification, one can be confident that an explanation is the best justified among possible alternatives. Experts are not seen as the ones who have simply gathered the most knowledge, or who assert their opinions the most persuasively. They are the ones who have used sound methods to generate evidence and provide reasonable interpretations. It is this evaluativist perspective which is seen as containing the understandings that allow for competent theory-evidence coordination.

Although there are no clear age ranges in which people obtain particular epistemic understandings, it is generally expected that absolutism is found more among schoolchildren, multiplism among middle adolescents, with some tendency toward a turn toward evaluativism in later adolescence and adulthood, particularly among those with higher education (Kuhn et al., 2000). However, in most age-group studies, all of the perspectives appear (Kuhn et al., 2000; Chandler et al., 2002; Barzilai and Weinstock, 2015). It has been argued that whereas the shift to understanding that there are multiple possible perspectives and relativism in knowing is a key developmental attainment (Chandler et al., 1990; Barzilai and Weinstock, 2020) there are several pathways that might follow this attainment including each of the general perspectives described above.

With that said, where does that leave research into the epistemic thinking of young school-age children? Even given the lack of clear age ranges in epistemic perspective development,

there is some evidence (Kuhn et al., 2000) and little reason to believe that young school-age children will have made a turn toward the types of knowledge relativism represented by the multiplist and evaluativist perspectives. Indeed, ToM research has found that around the ages of 4–5, when children recognize that people might hold different beliefs, they still hold that the beliefs come from objectively gained information and that only one belief is true and others are false (Carpendale and Chandler, 1996; Wellman et al., 2011). Thus we expect that young school-age children will essentially be absolutist, believing in the objectivity and possible certainty of knowledge. Epistemic thinking researchers have referred to attainment of ToM, and its recognition that there may be more than one knowledge claim, as an important conceptual precursor to the epistemic development that is described above (Burr and Hofer, 2002; Kuhn and Weinstock, 2002; Iordanou, 2016). However, research has not shown the transformation of this important marker of epistemic development into more multi-faceted personal theories of the nature of knowledge and knowing.

The current study is part of a research project which is based on the proposal that it is not false-beliefs ToM, but a later development, interpretive theory of mind (iToM) which is the important first step in the developmental task of trying to reconcile and coordinate objective and subjective aspects of knowing, the task that drives epistemic thinking development. It is with iToM that the notion of objective and subjective aspects first becomes a clear issue in knowing, so we expect to see the first hints of a turn toward personal theories of knowledge accompanying iToM attainment.

Children's attainment of theory of mind (ToM) is commonly assumed to be the watershed moment in understanding how people might assert different beliefs over the same matter of fact. In the false-beliefs task, 4–5-year-old children recognize that someone with incorrect information will form a different belief about an event than someone with correct information. For instance, a child with ToM who sees that a box of crayons actually contains candles will understand that someone who only sees the outside of the box will have a false belief about the contents of the box. (A child without ToM would claim the other person would know there were candles inside). Different claims are understood to arise when there are two sets of "facts," one associated with a real event and the other with a different event mistakenly assumed to be real. In this sense, a child might recognize that knowledge is relative to the knower, but only because people have different information, which the child understands has been obtained through objective processes (such as seeing). However, how do children understand how people may make different claims about an event even when they have the same information? With iToM, attained at about ages 7–8 (Lalonde and Chandler, 2002; Osterhaus et al., 2016), children understand that people might disagree because they interpret the same information differently. With this, it appears that children with iToM recognize that a claim need not be self-evident, but that information can be used as evidence for more than one claim, and knowledge claims have subjective sources in the knower's interpretation.

Specifically, we propose that ToM and the more advance iToM differ in how a child understands the knower and why people might assert different knowledge claims (Barzilai and Weinstock, 2020). ToM involves a type of perspective taking (about what someone knows or not given the information one has), and with the attainment of ToM children essentially hold that people know by using one cognitive function—perception—which they consider to be objective. In contrast, iToM does not involve taking a particular perspective but involves understanding that people do have subjective perspectives and make knowledge claims from these perspectives and may use a variety of cognitive activities (such as perception and interpretation). With other advanced, second-order ToM understandings children also understand that people gain knowledge through inference from information (Miller, 2009).

Although literature appears to point to attainment of simple ToM as a basis for children's understanding of NOS (Koerber et al., 2015), in fact studies apparently showing a relationship between ToM and reasoning skills relevant to NOS actually have samples with children of an age (8, compared with 4–5 when ToM is attained) in which it might be assumed that they have iToM or other more advanced theories of mind (Pillow, 1991; Carpendale and Chandler, 1996). Indeed, Koerber et al. (2015) do conclude that it seems that it is the children with more advanced theories of mind (at age 9) that demonstrate better understanding of NOS. The same researchers (Osterhaus et al., 2017) found that a measure of advanced ToM, which did not include iToM, did predict epistemic understanding of NOS.

In the current study, we explore how responses to the simple test of iToM—in which children with iToM claim that two people might have different interpretations of ambiguous pictures—might be represented in a more complex representation of interpretations of ambiguous evidence for scientific claims. Rather than just displaying an understanding that two people might produce different interpretations of the same picture, the children are asked to explain how two science experts might arrive at different conclusions about a biological phenomenon. As part of this research project, a preliminary quantitative analysis has found that indeed iToM, but not ToM, predicts epistemic understandings of the source, structure, and justification of knowledge that reflect the first grappling with subjective aspects of knowing within the absolutist perspective (Weinstock, 2018). The current analysis, however, is not concerned with tracking advancements in the epistemic perspectives of the developmental model (Kuhn and Weinstock, 2002). Rather, here we use a qualitative approach to describe the epistemic growth of children at two points in time looking at their expressions of issues such as perspectives and interpretation in the knowing process, the reasons for disagreement among experts, understandings of how knowledge is constructed and justified, and the nature of evidence. In addition to this developmental description, we assume that those demonstrating iToM at an earlier age, having understanding of subjective processes of knowing as a basis for knowledge claims, will show more advanced understandings of the epistemic issues across the 2 years of the study.

MATERIALS AND METHODS

Participants and Selection Criteria

This study sample was selected from a larger sample of 91 2nd graders (M_{age} = 7.05, range = 7.42–8.75) from two urban elementary schools who participated in a longitudinal research study that also included kindergarteners and 1st graders. All of the children were assessed three times at 1-year intervals. The current analysis includes a selection of these participants from the first and last years of the study, that is when they were in 2nd grade and then again from when they were in 4th grade. On the basis of responses to the assessments (to be described in the results section), the data of 28 2nd graders were included in the first wave and of 11 of these children when they were in 4th grade.

The population from which the sample was drawn was chosen to represent average Israeli children: The city is rated at the higher end of middle class (Central Bureau of Statistics, 2017), and each school has essentially average scores on a system-wide standardized test of academic skills (Ministry of Education, 2013). Among countries, Israel is ranked as having very high development (Human Development Report Office, 2015).

The research was approved by the Ministry of Education and the university human subjects research committee. The parents of each of the participants consented to the participation of their child. In addition, on the day a child was scheduled for the research, the child first gave consent to leave the classroom and then signed a consent form to participate in the research that was read aloud by a research assistant.

Materials

In each wave of the study, the children were assessed for iToM. They also responded each year to an epistemic perspective assessment in which two experts disagree about a biological phenomenon. This was immediately followed up by a semi-structured interview prompting reflection about the nature of knowledge and knowing, the reasons people might disagree about knowledge, and the components of knowledge justification. The semi-structured interview was the basis for the current, primarily qualitative analysis. Nevertheless, we describe the iToM task because we used this to select the sample for the current analysis as we explain in the results section. We also describe the epistemic perspective assessment because participants sometimes referred to it in their responses to the interview, so it is important to provide the context. All of the tasks and interviews were presented in one-on-one sessions with the researchers or research assistants in separate rooms. The sessions lasted between 30–45 min.

Interpretative Theory of Mind (iToM) Task (Lalonde and Chandler, 2002)

The participants were given the "droodles" task (Lalonde and Chandler, 2002) to assess for ToM and IToM. The task involves presenting a child with a picture and having the child describe the content of the picture. The child is then shown the same picture but framed in such a way that only an ambiguous portion of the picture is visible. The child is then asked to state what two different people (represented by adult-looking dolls) would say the picture they see is of. A child who states that the first doll would say that the picture is of what the child originally saw when looking at the whole picture is not considered to have achieved ToM (as this is essentially a false-beliefs task). A child who states that the first doll would say that the picture is something different from the original, whole picture would be considered to have passed the false-belief ToM task, as that child understands that the doll has different information than she or he does. But only if the child then states that the second doll sees something different from the first doll would the child be considered to have iToM in that the child recognizes that two people might have different interpretations of the same information. A child without iToM would insist that the two dolls must see, and know, the same thing as each other even if they do not know what is in the full picture.

There were four pictures. A point was given for each picture that a participant gave an iToM response to. Thus, each participant had an iToM score of 0–4.

Epistemic Perspective Assessment

An epistemic perspective assessment was developed for use with children at this age based on the format of discrepant accounts assessments used with older ages (Kuhn and Weinstock, 2002). The assessment is in the form of a picture book that tells the story of a new creature on a foreign planet. The prince/princess of the planet wants to understand what the creature uses its unusual hands for in order to explain the phenomenon to the king who has not seen the creature. Two expert advisors give conflicting explanations. One says the creatures has large hands with sharp nails to dig for food, and the other says these types of hands are used to hang on to tree branches when sleeping at night. During the course of the story, the participants were asked why the two experts did not agree, what they would do to decide on an explanation of the phenomenon, and how someone could decide about the correctness of a third person's claim about the phenomenon.

Analysis of responses to questions asked during the reading allowed for coding in levels that were conceptually consistent with three types of absolutism (realist, pre-dualist, and dualist), multiplism, and evaluativism (see Weinstock and Cronin, 2003) in three epistemic dimensions of source, justification, and structure (simplicity/certainty) of knowledge (Hofer and Pintrich, 1997).

Reflective Interview

The epistemic perspective was immediately followed by the more general questions:

(1) Is it possible for two people to see exactly the same thing and think different things from each other?
(2) Can you explain to me what knowledge is? What does it mean to know something?
(3) How do we know things about life?
(4) What is the difference between knowing and guessing?
(5) If you do not know something, what can you do to know it?

Although just a limited number of interviews from the total sample were used for the current analysis, all the participants responded to this interview. In order to understand the response

and develop themes, we read a larger number of interview responses than were ultimately used in the analysis. The coding scheme for the reflective interview was developed and employed with a series of iterative, collaborative processes. The goal of the researchers was to develop shared, describable representations of themes across interviews and categories of thinking within those themes.

One of the project researchers selected 10 of the interviews for use in developing the coding system which were ultimately not used in the final coding. The participants were selected to represent a range of iToM scores. The researcher who selected the interviews was not involved in their coding, and the two researchers developing the coding system were blind to the iToM scores.

As a first step in developing the coding system, the two researchers randomly selected five of the 10 interviews. They read each interview in full and, together, worked to make sense of the conceptions about knowledge and the processes of gaining knowledge that were expressed by the interviewee. After discussion, the researchers outlined the epistemic conceptions presented by each interviewee. Next, the researchers sought connections between conceptions of different interviewees on the same topic. In this way, the researchers defined themes that seemed to be repeated across interviews.

In the second step, the researchers worked separately to code the remaining five interviews based on the themes identified in the first five interviews. In this phase, the researchers' coding process was deductive, in analyzing the interviews based on the themes developed in the first step, and inductive in identifying expressions that extended the existing themes or suggested new themes. After coding three of these five interviews, the researchers compared the codes and resolved any differences or open questions. After comparing their separate analyses of the remaining two interviews, the researchers found that their codes were mostly similar and they referred to the same statements in support of the themes.

With this level of confidence, the coding system was formalized with a table of themes which were to be used to make a first pass at analyzing the 28 interviews from the 1st year of the study used in the current analysis. The 28 interviews were selected by a researcher not involved in the coding based on iToM scores. Fourteen of the participants had the very highest number of iToM responses, and the other 14 had the very lowest number of iToM responses. It should be noted that those developing the themes and coding these interviews according to them did not know which participants were in the high and low groups.

In this phase, the researchers elaborated detailed definitions for the themes while establishing them with quotations from the children. The inductive search for additional themes was continued throughout the coding, but the researchers verified that this system did indeed cover and exhaust the codes afforded by the data from the wide range of participants. This indicated that the coders had achieved inductive thematic saturation (Saunders et al., 2018) in developing the coding scheme.

At this point, the researchers began the deductive coding process of the 28 reflective interviews. This process was also undertaken in stages. First 10 of the interviews were randomly selected for coding by each of the researchers. They compared the coding of each interview and documented instances in which there were disagreements about how certain types of expressions would be coded. These disagreements were resolved and the documentation allowed for more elaborated and differentiated definitions of each theme. We followed the approach of achieving reliability through discussion and consensus in order to maximize the refinement of understanding of the coded ideas (Barbour, 2001; Campbell et al., 2013; Gläser-Zikuda et al., 2020). This process was repeated until all of the interviews were coded.

Whereas eight themes were developed, the current research focuses on two of them: (a) the dawning understanding of subjectivity as an explanation for how disagreements about knowledge arise, and (b) how one gains knowledge. The four different response types of how disagreements arise are described briefly here with coding examples given in **Table 1**. (1) Some children did not recognize or ignored discrepancies between the accounts of the scientists in the epistemic perspective assessment, and their discussion of the possibility of discrepant claims reflected this. They insisted there was and could only be a single, objective claim. Whereas most participants recognized discrepancies and acknowledged that people might

TABLE 1 | Category coding for the theme of explanations of how disagreements about knowledge arise.

Category	Characteristics	Examples
Single, objective claim	Competing claims not recognized as possible or not acknowledged.	Gave simple "no" in response to probe: "Is it possible for two people to see exactly the same thing and think different things from each other?" No spontaneous reference to competing claims in the remainder of the interview.
Different claims, no reason given	That people might make different claims is acknowledged but without any clear explanation or elaboration of why such disagreements might arise.	"It came out that way, because it might have come out that [it is different]."
People are different, having different minds with different content	Discrepancies explained by noting that people are different. They think differently because they are different. No elaboration on why people being different with different minds might lead to holding different knowledge.	"Let's say there's a fire somewhere, they'll think, one would think it's because there was too much heat, and one might think somebody lit a fire there and it spread. Everyone sees it differently." "Everyone thinks what he knows and thinks."
People have different perspectives, ways of thinking	That people have different lived lives, perspectives, and ways of knowing is the source of how they might arrive at different claims.	"It may be that Rona lives in a very large house and that is why this house looks small to her and Karen lives in a small house and now it looks big to her, so they can argue, because everyone thinks something different"

claim different things even if they had the same information, (2) some acknowledged this without providing any explanation, (3) some explained this as simply because people are different so their minds have different content, and (4) some offered reasons concerning different perspectives on information or different ways of thinking. It should be noted that in the original coding scheme there was an additional response, categorized between types 2 and 3. It was characterized as acknowledging differences with the explanation that people have different perceptual experiences. However, this response was seen in just one of the initial protocols used to develop the coding scheme, and in none of the protocols of the 28 participants analyzed for this study.

The response types are presented in order from objective to subjective, and within the subjective from less to more elaborated in explaining the source of the disagreements in terms of subjective processes of the knower. Ultimately, for the sake of comparison in the analysis, each participant was assigned to a response category based on their highest level of explanation.

There were three response types regarding how one gains knowledge (see **Table 2** for coding examples). Essentially, there were responses that referred to objective sources of knowledge—(1) perception or (2) external authority—and responses that referred to (3) aspects of individual knowledge construction, particularly investigation, although not much to interpretation, as the source of knowledge. For the sake of comparison in the analysis, participants were assigned to response types according to dominant answer. None of those assigned to response types 1 or 2 (i.e., both representing knowledge as gained passively through external sources) gave responses consistent with type 3 (knowledge gained through active construction).

TABLE 2 | Category coding for theme of understandings of how knowledge is gained.

Category	Characteristics	Examples
Senses, direct perception	Knowledge gained through the senses	"If you know something, you must have listened to something or seen."
External authority	Knowledge is learned, passively acquired from authoritative external source, presenting information, such as teachers, parents, friends, books, or the internet.	". . .learn, learn. from school or high school or anywhere."
Investigation, exploration	Knowing involves intentional activity on the part of the knower. Goal-directed investigation, exploration, and questioning are needed to confirm and develop ideas. Might accept others as ultimate objective authorities, but knowing requires effort on the part of the knower.	"Knowledge is when you research something and you know it really already, without mistakes." "There are many methods. Go and ask, do research. . ..Come to class and say I am in kindergarten and I want to know what it is like to be in first grade and then you are told. You ask a child to bring all the school-related things: a symbol, a set of hours, everything and then he gives an explanation."

Certain themes, such as these two, emerged clearly in the words of all the children in response to questions in the interview. Because they were easy to identify in all the children these themes suggested a conceptual sequence and it was possible to place each child's responses in that sequence. Moreover, understanding the possibility of multiple, disagreeing perspectives is seen as a foundational aspect of epistemic development (Kuhn et al., 2000; Barzilai and Weinstock, 2020), as is understanding how one gains knowledge (Miller et al., 2003; Pillow, 2008; Fitneva et al., 2013).

The other themes (checking the truth of a knowledge claim, general understanding of the sources of knowledge, conceptions of expertise, conceptions of absolute truth, certainty, and the difference between guessing and knowing) are also informative regarding the epistemic thinking of children; that they were expressed spontaneously in some, but not all participants, shows that at least some children at this age attended to such issues. However, for the sake of the current longitudinal analysis they will remain in the background. These themes tended to represent specific issues subsumed in the broader themes of why there are disagreements and how people come to know. Moreover, several of these themes, such as checking the truth of knowledge claims, were not expressed by all of the interviewees. On one hand, the lack of expression of these themes may be meaningful in that this could indicate that the children had not yet developed concepts regarding such issues (i.e., the children may not have considered that knowledge claims need to be verified), or no longer held the concepts (i.e., that knowledge is absolute). On the other hand, it is also possible that the interview questions simply did not sufficiently probe these issues. In either case, these themes could not be found consistently expressed by all participants in the interviews conducted in the first and last years.

RESULTS AND DISCUSSION

As mentioned above, the responses to 28 semi-structured interviews from grade 2 were chosen according to level of iToM exhibited by the children. Because of the demands of coding the interviews, and because the major goal of the analysis is to describe the range of epistemic development rather than focus on hypothesis testing, the interviews were selected according to criteria rather than randomly to try to capture a range of epistemic thinking. The selection criteria at grade 2 and grade 4 were used in order to be able to develop a comprehensive picture of epistemic development over 2 years leading from iToM attainment. The reflective interviews from grade 2 chosen for analysis included those from participants with the highest (n = 14) and lowest (n = 14) number of iToM responses. Eleven of those reflective interviews were selected for comparison with those participants' reflective interviews from 4th grade. The 11 were selected according to responses they had made in 2nd grade so that the development from a range of responses could be described across the 2 years. The interviews of these 11 participants were used for the longitudinal comparative analysis between 2nd and 4th grade presented here. The themes developed at the first wave were further elaborated and refined for this analysis by one of the researchers who had been involved in

developing and employing the coding scheme in the first wave. The refinements included expressions that were found in the discussions in developing the themes with the grade 2 interviews but were not frequent enough to warrant distinction from the categories found. At grade 4, with greater salience, it became possible to make more specific distinctions.

The interviews from 2nd grade were analyzed to identify overall epistemic themes. As described earlier, two themes were identified consistently across participants: understanding why disagreements arise about knowledge claims and understandings of how knowledge claims are constructed and justified. The first theme has particular importance because it reflects the consideration of subjective perspective in knowing which is the issue which we assume is the essential proposed connection between iToM and further epistemic development. We used this classification of responses in describing the participants in the comparison between 2nd and 4th grade as we were interested in tracking shifts toward relativist thinking. Before analyzing the comparison between the grades we present the distribution of participants across response types in these themes from 2nd grade in order to illustrate their relationship to iToM attainment.

2nd Grade Interviews

There were four types of responses reflecting understanding of the source of discrepant claims that the participants gave, primarily to the interview question, "Is it possible for two people to see exactly the same thing and think different things from each other?" However, other places in the interview where the participant mentioned the disagreement between the scientists in the story were also taken into consideration. The response types are listed in **Table 3** in order of our assumptions about epistemic development (Kuhn and Weinstock, 2002). (Refer to **Table 1** for the definitions and examples of each response type). **Table 3** shows the distribution of these response types by high and low iToM response. Half of the analyzed sample was chosen because they gave no or the fewest number of iToM responses in the iToM task while the other half had given the highest number of iToM responses.

As can be seen in **Table 3**, those with low iToM tended much more than those with high iToM to state that there was one objective claim that was correct or did not give reasons why the scientists had given different claims. In contrast, those with high iToM almost entirely focused on the subjectivity of the scientists, saying they were different people with different ideas or that they had reasons or backgrounds to take different perspectives in making their claims. Just to confirm this pattern, a 2 × 2 Fisher's exact text, which can be used with small samples, was performed with not recognizing or not finding reasons for discrepant claims in one group, and those acknowledging subjective states or processes being in the other group. The test was significant, $p = 0.038$.

Table 4 shows the distribution of response types regarding understandings of how people gain knowledge. (Refer to **Table 2** for definitions and coding examples). These responses were to the interview question, "If you do not know something, what can you do to know it?" However, other places in the interview where the participant mentioned the source of knowledge were also taken into consideration. As can be seen in the table, almost all with low iToM gave one of the objectivist responses. Those with high iToM were split among the objective and constructive response types, but almost all of those expressing constructivist ideas were those with high iToM. Again, a 2 × 2 Fisher's exact test showed an association between level of iToM and response type, $p = 0.038$.

TABLE 3 | Explanations of how disagreements arise about knowledge by iToM response level.

Discrepant claims explanation	iToM response level	
	Low iToM (n = 14)	High iToM (n = 14)
Single, objective claim	1	0
Different claims, no reason given	5	1
People are different, different minds with different content	7	10
People have different perspectives, ways of thinking	1	3

2nd-4th Grade Comparisons

For the comparative analysis between 2nd and 4th grade, the intent was to choose the interviews from both grades of four children from each the "different claims, no reason," "people are different/have different minds," and "people have different perspectives/ways of thinking" categories to examine the trajectories of development from those starting places. As one of the children from the "different perspectives" category had dropped out of the study by the last year, and there were only three left in this group, the total sample was 11.

Understanding of Disagreements Between Knowledge Claims

For the issue of understanding of how disagreements about knowledge can arise, the participants' responses at grades 2 and 4 are presented here in pairs in order to illustrate the development. The cases we are presenting were chosen to be representative of the refined, more specific themes developed in this analysis. For each participant, the iToM categorization and the categorization for understanding of sources of disagreement in 2nd grade are noted. The understanding disagreeing claims categorization in 4th grade is also noted. The first finding, which will be illustrated, is that the different perspectives and ways of thinking category was differentiated. Some responses noted that there were different perspectives and ways of thinking while another response type, which is apparently more advanced, specified internal thinking

TABLE 4 | Understandings of how knowledge is gained by iToM response level.

Knowledge sources	iToM response level	
	Low iToM (n = 14)	High iToM (n = 14)
Senses, direct perception	2	1
External authority	11	7
Investigation	1	6

processes as the source of individual's knowing that could result in disagreements. Although we do not present all of the 11 cases, it should be noted that in all cases, except for those who had already expressed understanding of internal processes in 2nd grade, there were epistemic shifts toward more consideration of the subjective aspects of knowing from grade 2 to grade 4.

The first comparison includes two cases of participants who had low iToM in 2nd grade, and who were responding to the question, "Is it possible for two people to see exactly the same thing and think different things from each other? Why/why not?" In each case, in 2nd grade the participants imply that subjective differences can be overcome so disagreement is not necessary (coded as different claims noted, but giving no reasons to explain this). By 4th grade each acknowledges that subjectivity is at the root of knowing, but with a focus only on opinion or motivation with no mention of interpreting information differently (coded as different perspectives or ways of thinking). For instance, as a 2nd grader P116 states: "Because it's not the same people and they do not know each other" with the implication that if they did know each other they would not have disagreed or they could work out their disagreement. In either case, there is no explanation of why not knowing each other would lead to disagreement between people. In 4th grade, P116 states: "Everyone has her own opinion" giving a reason for the disagreement and focusing on subjectivity. The stress on opinion is the hallmark of multiplism, the most subjective epistemic perspective (Kuhn and Weinstock, 2002).

Similarly, 2nd grader P74 says: "Because he doesn't not need to know what I know, If I tell him then he would know." Although approaching the different people have different minds with different content category, in this case the participant diminishes the effect of this and does not explicitly explain how this should lead to different claims. The tone of P74's statement is that one person knows and the other does not, but could know. It is not that the minds have different content, but one mind has more information than the other. In response to a question from the epistemic perspective assessment about how it could be that the expert royal advisors had two different answers, 2nd grader P74 said, "Because they did not hear each other they think differently." That is, as long as people manage to have the same information, they will agree. At 4th grade P74 does give an explanation for differences rooted in subjectivity, "Because we are not the same person, if we were thinking the same thing it would be boring." But, in response to the epistemic perspective assessment questions, in 4th grade P74 still thinks that subjectivity is just a distraction and that ultimately different knowledge claims can be adjudicated by sharing information: "In the end they will reach a decision and they both will be right. They'll say: Let's go and see."

Similar to the opinion and motivation oriented reasons given in 4th grade by P116 and P74, a participant (P249) who had high iToM in 2nd grade said this in 2nd grade: "Everyone has their own opinion, not everyone has the same opinion." By 4th grade, this participant focuses in a range of internal processes and bases for multiple perspectives, of which opinion is just one:

"There can be a situation like this, as we learned about the two characters who managed to discover two different things from the same part of a picture [from the iToM task]. Everyone can hear a different side and another place, another story and another opinion."

Rather than just throwing out the notion that people have different perspectives, P249 specifies what perspective might mean and hints at the implications of people holding different perspectives. P144, who also was high iToM and started out in 2nd grade straightforwardly expressing the different ways of thinking response, although with no mention of opinion—"Everyone thinks differently"—also further specified the internal processes of thinking that could produce discrepant claims. In 4th grade he said, "Everyone has his imagination, his world, and thinks differently."

From the examples given so far, we can see that some of the participants seemed to be working out the role of subjectivity in knowing. This trend started with the acknowledgment of subjectivity without any clear explanation of its necessity or why it played a role. This advanced to the unavoidability of subjectivity, especially with reference to opinion. Then a few specified subjective ways of thinking beyond opinion, but with a focus on idiosyncratic characteristics as the source of perspectives.

In another trend, the development moves toward consideration of processes of interpreting and emphasizing information and is more empirically oriented. That is, there is more of a focus on knowledge issues than character issues, although character issues still come into play. There are also explicit and implicit references to how fields of expertise or interest influence ways of thinking. This can be seen in the responses of two participants who were high iToM and were categorized as focusing on internal processes as the reason for disagreement at both grades 2 and 4. In 2nd grade, P151 expresses internal processes somewhat in line with the emphasis on idiosyncratic characteristics and preferences: "There is a Lego that one likes and another doesn't, so everyone says something different about it." In 4th grade, there is a turn from preferences to attributing disagreements to fields of expertise and the subsequent effect on thinking: "Because they seem to be experts on different things so they will think differently." 2nd grader P180 starts with a seeming reference to expertise and its relation to ways of thinking, but defining it as a matter of preference. She said:

"There are people who like tree tops and some who love plants so maybe they explain different things because they think that way."

She still focuses on preference in 4th grade, but unlike the previous trend she mentions empirical implications.

"Everyone has a different character and a different look, and each one sees from one's angle. One will see fingernails on the little finger and the other on the thumb."

With this, there is developing appreciation of expertise and perspective that it brings. In response to the question from the epistemic perspective assessment about how it could be that the expert royal advisors had two different answers, 2nd grader P248, who had low iToM and did not give reasons to explain why people might disagree, simply said: "Because neither of them knows." By 4th grade, he said, "Both are correct. One is an expert on trees and one is an expert on land." This is consistent with

his response to the reflective interview in which he said that disagreement exists because people have different perspectives and ways of thinking. People might disagree not just because of idiosyncratic perspective and opinion, but because of emphases in what they know and how they explain given their field of expertise. P249, who had high iToM had a similar response to the same question, but in 2nd grade. "I think because each one is an expert on something else and each thinks of something she is an expert in." In 4th grade, P249 expresses how perspective complicates the process of knowing. "That implies one does not always come to an exact answer. This is problematic." This view, that some uncertainty and tentativeness is part of knowing, is decidedly not absolutist.

Understandings of How Knowledge Is Gained

The children in this study had a harder time expressing what they think of how knowledge claims are constructed and justified, even in 4th grade. This is presaged by the coding of their understandings of how knowledge is gained from their 2nd grade interviews as outlined in **Table 4**. Just one person with low iToM referred to investigation as a way to gain knowledge, as did minority of even those with high iToM. In either 2nd or 4th grade, those with high iToM did tend to indicate the need to develop a base of evidence in response to the reflective interview questions (i.e., "Can you explain to me what knowledge is? What does it mean to know something?", "How do we know things about life?", "What is the difference between knowing and guessing?", and "If you do not know something, what can you do to know it?"). However, as will be seen, their ideas of what counts as good evidence is not always clear, although in contrast to those with low iToM they did indicate the need for evidence. The examples given only include those with high iToM, because those with low iToM simply did not refer to evidence. (One notable characteristic was that several told fantastical stories about the creature that were not even based on the information about the creature in the story book).

To give an example of what developed, in response to the question about the difference between knowing and guessing, P249 gave this answer as a 2nd grader in the first year of the study:

"When I guess something, I'm not sure, but a guess a hypothesis. 'Knowledge' – if you're still learning, you still do not know everything. There are all kinds of answers, and there's a chance you'll answer that you did not learn about it and you will not know the correct answer and you will be wrong."

Whereas this participant suggests that multiple answers must be considered, it seems that the participant regards knowledge as being mostly certain whereas a guess is not correct and is wrong. As was expected with iToM, there is a suggestion of subjective processes in knowing, in particular in coming up with the wrong answer. However, there is no mention of evidence. But in the third year of the study when the participant is a 4th grader, the participant does see that knowing requires some type of evidence.

"To know something is to be certain of something. that you have evidence, and you saw it in your eyes. A guess is a hypothesis of something you cannot be sure of and you think about."

Although this participant does say one needs evidence, the idea of evidence remains at the level of personal observation. Other participants with high iToM offered more complex views of evidence. For instance, in response to the same question, P151 said as a 4th grader in year 3 of the study:

"He is going to see, and was confirmed by several witnesses. Everyone comes because he can lie. . .. Anyone can imagine differently (from the other) and then check who was right."

Although not directly talking about justification, this participant does say that claims need to be supported by evidence, and that reports of evidence and claims need to be judged and justified.

Also in response to the question regarding the difference between knowing and guessing, P144 said:

"Knowing that you know more about this and that is pure. You saw, you read, you heard. To guess, you just think about it with your head."

In this, the participant mentions that to know one needs multiple sources of evidence, and notably, not just evidence based in direct observation. We also see here indications of recognizing that subjectivity plays a role in claims, again particularly in those that are less justified.

Finally, two participants when in 2nd grade study gave responses that seem to recognize the need to gather and provide information from different sources. Although some of their suggestions are not likely realistic forms of evidence (getting the answers from the creature), and other suggestions mention what might be good sources without specifically mentioning the evidence they would get, it is clear that they know that claims are not self-evident and information must be generated to justify a claim. For example, P99 said:

"I can take the creature to the house, to the kingdom and then I can test it, the hands, maybe if he speaks our language – I can also ask him. And just bring it and look at his hands. You can also go outside sometimes and look at what he's doing with them."

P141 said:

"I'll try to talk to him and ask him', Where did you come from? Are you a monster? Why did you come here?' . . . I think maybe because he has such pink red spots he eats raspberries. . . [says to the experimenter] Go back to the picture of the creature. . . Yes, he eats things from the tree. And I think he's green because he's eating cucumbers and green things like lettuce. . . I'll look for him and ask him all these questions. And if he does not know how to speak, then I'll call my advisors to tell me how they understand him when he talks, that's all."

Although some of this participant's comments might be seen to be speculative, based on limited observation, and off the topic of knowing why the creature has big hands and sharp claws, the participant offers evidence from claims for the creature's coloring and suggest different acceptable sources of evidence, the creature's testimony, observation, and the experts' testimony.

Although none of the examples above show much understanding of either justification or evidence, it should be noted that the participants were not asked directly to produce either. What are seen are inklings of understanding that information and reliable sources need to be offered to justify claims to oneself and others. This stands in contrast to those with low iToM not quoted who offered no justification, repeated the explanations of the experts, or gave made up,

irrelevant information or means of knowing (e.g., "you could use a magic mirror").

GENERAL DISCUSSION

The longitudinal analysis of qualitative interviews provides support for two of the assumptions tested in the larger research project. For one, it has been assumed that the underlying task of epistemic development is coordinating the objective and subjective aspects of knowing. For this task to be approached, someone must appreciate the existence of and come to identify what might be subjective in knowing. In this research, we see that children are beginning this work. While some of the children at first, in 2nd grade, did not recognize conflicts in claims or noted them without being able to explain them, 2 years later children seem to have started work on this issue. By 4th grade, at the age of 9–10, all suggested that the difference in claims had some subjective reason, at least in differences of perspective and general ways of thinking if not in more specific internal cognitive or motivational processes. Whereas in 2nd grade, some suggested differences might be attributed to lack of knowledge, which could be easily remedied, by 4th grade the participants were pointing more toward differences between people.

Interestingly, at this level some of the participants emphasized more the subjective side of things, such as focusing on opinion or preference, while others hewed more closely to the objective side considering differences in emphasis on information with expertise. This might have been because at this age, children might still have absolutist views that ultimately one can find a single correct, objective answer. With this, the results provide some description of how children consider subjectivist aspects of knowing at the absolutist level which then may lead to the transition to multiplism. Alternatively, perhaps this difference is a precursor of what is seen in later development, such as with different types of multiplism with some holding that all perspectives are hopefully idiosyncratic and others holding that claims can be justified, but the multiplicity of perspectives is an impediment (Perry, 1970; Weinstock and Cronin, 2003). Also, in evaluativism there are objective relativists, for whom perspective differences lie with methodological differences and emphases on different evidence, and conceptual relativists, for whom perspective differences lie with systematic subjective contexts and individual interpretations (Leadbeater and Kuhn, 1989). This split in emphases on the objective or subjective aspects of knowing at different levels of epistemic development suggests it might be worthwhile to investigate cognitive, cultural, or individual factors in people's orientations toward objectivity or subjectivity.

The second assumption of this research project supported by the results is that the development of iToM is significant in the turn toward grappling with subjectivism in knowing, and thus, toward epistemic development toward relativism. This helps provide a more complete picture of the development of epistemic development through the lifespan and provides a link between development of ToM and continuing developments of theories of knowledge. However, although the meaning of iToM is that children recognize that different knowledge claims arise because people interpret information differently, there was little evidence in the children's interviews that they attributed claim differences between the scientists or in general to interpretation. Perhaps iToM sets children on the course to recognized subjective differences in how people receive, process, and emphasize information but the understanding of how interpretations arise and are used in knowledge building only comes later in epistemic development. This would be consistent with Kuhn and Weinstock's (2002; see Barzilai and Weinstock, 2015) model of epistemic development which puts understanding that knowledge is interpretation at the heart of evaluativist thinking.

Whereas the research shows considerations of the meaning of perspective in knowing, it also points to fuzzy understanding of evidence and no real reference to how it relates to theoretical explanations. In some sense, this is not surprising given that young children struggle with differentiating evidence from explanation (Kuhn and Pearsall, 2000), and even adolescents who can distinguish between them find it easier to produce explanations (Glassner et al., 2005). However, the children in this study seemed more oriented toward the generation of evidence than toward producing explanations. Except for P180, who said that experts would explain things differently, they spoke little about explanation as something constructed. They seemed to expect that evidence and the gathering of information would point to one explanation or another. The search for evidence—such as talking to or directly observing the creature—would confirm the claim of one of the scientists. Good evidence is equated with first-hand observation. Thus, it seems that theory-evidence coordination is not supported by the epistemic understandings at this age.

In their seminal article on the understanding of the NOS, Carey and Smith (1993) distinguish between the two epistemologies of "knowledge unproblematic" and "knowledge problematic." Although most of the participants' expressions would seem to fall in the knowledge unproblematic category, that correct knowledge can be known certainly and that opinions might explain different claims, perhaps we do see hints of the progression toward the knowledge problematic. For instance, P180 in mentioning that how one thinks depends on one's knowledge, expertise, and "angle" when looking at something, suggests a nascent understanding of interpretive frameworks. P249, who in 2nd grade gives the knowledge unproblematic response that people disagreed because they had different opinions without explanation, by 4th grade explains that people might draw different conclusions from the same perceptual experiences, a characteristic of the knowledge problematic, in saying that people can "discover two different things from the same part of a picture." As if he had read Carey and Smith's article, P249 goes on to say, "That implies one does not always come to an exact answer. This is problematic."

It is likely that children, at least in 4th grade, do have a better understanding of how scientists investigate more than is apparent in the interviews, and that such understanding is based on epistemic understanding (Osterhaus et al., 2017). It is possible participants did not recognize the scientific issue in

this study in the same way that they understand formal science in a school context. Aside from the context, the presentation of science issues through discrepant claims, as opposed to as open or answered questions, is not the common way of talking about science in the academic context. In one study (Tabak et al., 2010), biology undergraduates reported that they had not been exposed to discrepant claims as part of their coursework, in contrast with history undergraduates who reported that working through discrepant claims was a central part of their coursework. In addition, it is not too surprising that children do not have a sense of how justification in science takes place, other than through observation, when adolescents and many adults also are not particularly capable in distinguishing between the quality of science arguments which may or may not have reflected the differentiation of theory and evidence (Barchfeld and Sodian, 2009).

The qualitative analysis exploring and describing aspects of epistemic development suggests hypotheses regarding epistemic and NOS development that could be tested with quantitative studies. One, which was referred to earlier, is an analysis underway investigating whether iToM development predicts dimensions of epistemic thinking regarding the source, structure, and justification of knowledge (Weinstock, 2018). This research can be further developed to look at developments found in the current study between the 2nd and 4th grade in understanding disagreements and how knowledge is gained are reflected in the dimensions of source, structure, and justification of knowledge. In specific epistemic understanding of NOS, research could look at whether the development of iToM is a factor in explaining the trajectory of development found in research looking at the understanding of experiments (Osterhaus et al., 2017), theorizing (Metz, 2004; Koerber et al., 2015), and the coordination of theory and evidence (Kuhn et al., 2008). The current study did not reveal great understanding of hypothesis testing or the distinction between hypothesis and evidence—for instance, in response to the question about the distinction between guessing and knowing, even among those with high iToM the term hypothesis was used in relation to the term guess. Nevertheless, it would be worthwhile to examine whether the early understanding of perspective and interpretation found in iToM would be related to whether claims made from different perspectives might be seen as hypotheses which could be tested and if there is understanding of how evidence could be generated and brought to bear in testing such hypotheses.

Whereas we have looked at epistemic understanding as an outcome of development, it is important to note that we are not proposing that epistemic understanding, and the consequent understanding of knowledge in science and other disciplines, will come as a matter of course. Education, beyond age, has been found to be a factor in epistemic development (King and Kitchener, 1994; Kuhn et al., 2000; Tabak et al., 2010). There has been little research on how formal education might influence ToM, and particular iToM development, but there is evidence that mothers' talk can promote earlier development of ToM (Ruffman et al., 2002) and iToM (Tafreshi and Racine, 2016), and that parents' expression of epistemic information and their epistemic perspectives impact on their children's evidence talk (Luce et al., 2013). Thus, how young children grapple with issues of epistemology and develop epistemic understandings, as have been found in this research, should help contribute to educational projects in the everyday understanding of science and how knowledge in general is constructed and justified.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Chief Scientist's Office, Israeli Ministry of Education and Ben-Gurion University of the Negev Human Subjects Research Committee. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin and agreed to by the child participant.

AUTHOR CONTRIBUTIONS

MW is the PI of the research project. He planned the research, led the development of all of the research materials, worked with each of his collaborators on the coding and analyses of the data, and was the lead author on the manuscript. VI led the qualitative analysis across the different timepoints of the study. This analysis and the manuscript are based in part on her MA thesis. HC led the qualitative analysis in the first year of the study. The analysis and the manuscript are based in part on her MA thesis. IT worked with VI and HC in developing the coding scheme and analysis of the interviews. YH led the implementation of the research. Along with the PI, she led the development of all the research materials. She led the coding of the interpretive theory of mind task.

FUNDING

This research was supported by a grant from the Israel Science Foundation (1570/12).

REFERENCES

Barbour, R. S. (2001). Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *BMJ* 322, 1115–1117. doi: 10.1136/bmj.322.7294.1115

Barchfeld, P., and Sodian, B. (2009). Differentiating theories from evidence: the development of argument evaluation abilities in adolescence and early adulthood. *Informal Logic* 29, 396–416.

Barzilai, S., and Weinstock, M. (2015). Measuring epistemic thinking within and across topics: a scenario-based approach.

Contemp. Educ. Psychol. 42, 141–158. doi: 10.1016/j.cedpsych.2015.06.006

Barzilai, S., and Weinstock, M. (2020). "Beyond trustworthiness: comprehending multiple source perspectives," in *Handbook of Learning From Multiple Representations and Perspectives*, eds P. Van Meter, A. List, D. Lombardi, and P. Kendeou (New York, NY: Routledge), 123–140. doi: 10.4324/9780429443961-11

Barzilai, S., and Zohar, A. (2012). Epistemic thinking in action: evaluating and integrating online sources. *Cogn. Instr.* 30, 39–85. doi: 10.1080/07370008.2011.636495

Burr, J. E., and Hofer, B. K. (2002). Personal epistemology and theory of mind: deciphering young children's beliefs about knowledge and knowing. *New Ideas Psychol.* 20, 199–224. doi: 10.1016/S0732-118X(02)00010-7

Campbell, J. L., Quincy, C., Osserman, J., and Pedersen, O. K. (2013). Coding in-depth semistructured interviews: problems of unitization and intercoder reliability and agreement. *Sociol. Methods Res.* 42, 294–320. doi: 10.1177/0049124113500475

Carey, S., and Smith, C. (1993). On understanding the nature of scientific knowledge. *Educ. Psychol.* 28, 235–251. doi: 10.1093/acprof:oso/9780195115772.003.0003

Carpendale, J. I. M., and Chandler, M. J. (1996). On the distinction between false belief understanding and subscribing to an interpretive theory of mind. *Child Dev.* 67, 1686–1706. doi: 10.2307/1131725

Central Bureau of Statistics (2017). *Characterization and Classification Of Geographical Units by the Socio-Economic Level Of the Population, 2013 (Publication No. 1694)*. Israel: Central Bureau of Statistics.

Chandler, M., Boyes, M., and Ball, L. (1990). Relativism and stations of epistemic doubt. *J. Exp. Child Psychol.* 50, 370–395. doi: 10.1016/0022-0965(90)90076-K

Chandler, M. J., Hallett, D., and Sokol, B. W. (2002). "Competing claims about competing knowledge claims," in *Personal Epistemology: The Psychology of Beliefs About Knowledge and Knowing*, eds B. K. Hofer and P. R. Pintrich (Mahwah, NJ: Erlbaum), 145–168.

Fitneva, S. A., Lam, N. H., and Dunfield, K. A. (2013). The development of children's information gathering: to look or to ask? *Dev. Psychol.* 49, 533–542. doi: 10.1037/a0031326

Gläser-Zikuda, M., Hagenauer, G., and Stephan, M. (2020). The potential of qualitative content analysis for empirical educational research. *Forum* 21:3443. doi: 10.17169/fqs-21.1.3443

Glassner, A., Weinstock, M., and Neuman, Y. (2005). Pupils' evaluation and generation of evidence and explanation in argumentation. *Br. J. Educ. Psychol.* 75, 105–118. doi: 10.1348/000709904X22278

Hofer, B. K., and Pintrich, P. R. (1997). The development of epistemological theories: beliefs about knowledge and knowing and their relation to learning. *Rev. Educ. Res.* 67, 88–140. doi: 10.3102/00346543067001088

Huang, Z., Hu, Q., and Shao, Y. (2019). Understanding others' knowledge certainty from inference and information-seeking behaviors in children. *Dev. Psychol.* 55, 1372–1379. doi: 10.1037/dev0000729

Human Development Report Office (2015). *Human Development Report 2015: Work for Human Development*. New York, NY: United Nations Development Programme.

Iordanou, K. (2010). Developing argument skills across scientific and social domains. *J. Cogn. Dev.* 11, 293–327. doi: 10.1080/15248372.2010.485335

Iordanou, K. (2016). From theory of mind to epistemic cognition. A Lifespan perspective. *Frontl. Learn. Res.* 4:106–119. doi: 10.14786/flr.v4i5.252

King, P. M., and Kitchener, K. S. (1994). *Developing Reflective Judgment: Understanding and Promoting Intellectual Growth and Critical Thinking in Adolescents and Adults*. San Francisco, CA: Jossey Bass.

Kloo, D., Rohwer, M., and Perner, J. (2017). Direct and indirect admission of ignorance by children. *J. Exp. Child Psychol.* 159, 279–295. doi: 10.1016/j.jecp.2017.02.014

Koerber, S., Osterhaus, C., and Sodian, B. (2015). Testing primary-school children's understanding of the nature of science. *Br. J. Dev. Psychol.* 33, 57–72. doi: 10.1111/bjdp.12067

Kuhn, D. (1991). *The Skills of Argument*. New York, NY: Cambridge University Press.

Kuhn, D. (2001). How do people know? *Psychol. Sci.* 12, 1–8. doi: 10.1111/1467-9280.00302

Kuhn, D., Cheney, R., and Weinstock, M. (2000). The development of epistemological understanding. *Cogn. Dev.* 15, 309–328. doi: 10.1016/S0885-2014(00)00030-7

Kuhn, D., Iordanou, K., Pease, M., and Wirkala, C. (2008). Beyond control of variables: what needs to develop to achieve skilled scientific thinking? *Cogn. Dev.* 23, 435–451. doi: 10.1016/j.cogdev.2008.09.006

Kuhn, D., and Pearsall, S. (2000). Developmental origins of scientific thinking. *J. Cogn. Dev.* 1, 113–129. doi: 10.1207/S15327647JCD0101N_11

Kuhn, D., Wang, Y., and Li, H. (2010). Why argue? Developing understanding of the purposes and values of argumentive discourse. *Discourse Process.* 48, 26–49. doi: 10.1080/01638531003653344

Kuhn, D., and Weinstock, M. (2002). "What is epistemological thinking and why does it matter?," in *Personal Epistemology: The Psychology of Beliefs About Knowledge and Knowing*, eds B. K. Hofer and P. R. Pintrich (Mahwah, NJ: Erlbaum), 121–144.

Lalonde, C. E., and Chandler, M. J. (2002). Children's understanding of interpretation. *New Ideas Psychol.* 20, 163–198. doi: 10.1016/S0732-118X(02)00007-7

Leadbeater, B., and Kuhn, D. (1989). "Interpreting discrepant narratives: hermeneutics and adult cognition," in *Everyday Problem Solving*, ed. J. Sinnott (New York, NY: Praeger), 175–190.

Luce, M. R., Callanan, M. A., and Smilovic, S. (2013). Links between parents' epistemological stance and children's evidence talk. *Dev. Psychol.* 49, 454–461. doi: 10.1037/a0031249

Mason, L., Gava, M., and Boldrin, A. (2008). On warm conceptual change: the interplay of text, epistemological beliefs, and topic interest. *J. Educ. Psychol.* 100, 291–309. doi: 10.1037/0022-0663.100.2.291

Mason, L., and Scirica, F. (2006). Prediction of students' argumentation skills about controversial topics by epistemological understanding. *Learn. Instr.* 16, 492–509. doi: 10.1016/j.learninstruc.2006.09.007

Metz, K. E. (2004). Children's understanding of scientific inquiry: their conceptualization of uncertainty in investigations of their own design. *Cogn. Instr.* 22, 219–290. doi: 10.1207/s1532690xci2202_3

Miller, S. A. (2009). Children's understanding of second-order mental states. *Psychol. Bull.* 135, 749–773. doi: 10.1037/a0016854

Miller, S. A., Hardin, C. A., and Montgomery, D. E. (2003). Young children's understanding of the conditions for knowledge acquisition. *J. Cogn. Dev.* 4, 325–356. doi: 10.1207/S15327647JCD0403_05

Ministry of Education (2013). *Meitzav [Measures of School Efficiency and Growth]. [in Hebrew]*. New Delhi: Ministry of Education.

Nussbaum, E. M., Sinatra, G. M., and Poliquin, A. (2008). Role of epistemic beliefs and scientific argumentation in science learning. *Int. J. Sci. Educ.* 30, 1977–1999. doi: 10.1080/09500690701545919

Osterhaus, C., Koerber, S., and Sodian, B. (2016). Scaling of advanced theory-of-mind tasks. *Child Dev.* 87, 1971–1991. doi: 10.1111/cdev.12566

Osterhaus, C., Koerber, S., and Sodian, B. (2017). Scientific thinking in elementary school: children's social cognition and their epistemological understanding promote experimentation skills. *Dev. Psychol.* 53, 450–462. doi: 10.1037/dev0000260

Perry, W. G. Jr. (1970). *Forms of Intellectual and Ethical Development in the College Years: A Scheme*. New York, NY: Holt, Rinehart, and Winston.

Pillow, B. H. (1991). Children's understanding of biased social cognition. *Dev. Psychol.* 27, 539–551. doi: 10.1037/0012-1649.27.4.539

Pillow, B. H. (2002). Children's and adults' evaluation of the certainty of deductive inferences, inductive inferences, and guesses. *Child Dev.* 73, 779–792. doi: 10.1111/1467-8624.00438

Pillow, B. H. (2008). Development of children's understanding of cognitive activities. *J. Genet. Psychol.* 169, 297–321. doi: 10.3200/GNTP.169.4.297-321

Ruffman, T., Slade, L., and Crowe, E. (2002). The relation between children's and mothers' mental state language and theory-of-mind understanding. *Child Dev.* 73, 734–751. doi: 10.1111/1467-8624.00435

Ryu, S., and Sandoval, W. A. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. *Sci. Ed.* 96, 488–526. doi: 10.1002/sce.21006

Sandoval, W. A., and Çam, A. (2011). Elementary children's judgments of the epistemic status of sources of justification. *Sci. Ed.* 95, 383–408. doi: 10.1002/sce.20426

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., et al. (2018). Saturation in qualitative research: exploring its conceptualization and operationalization. *Qual. Quan.* 52, 1893–1907. doi: 10.1007/s11135-017-0574-8

Sodian, B., Zaitchik, D., and Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Dev.* 62, 753-766. doi: 10.1111/j.1467-8624.1991.tb01567.x

Tabak, I., Weinstock, M., and Zviling-Beiser. (2010). "Discipline-specific socialization: a comparative study," in *Proceedings of the 9th International Conference of the Learning Sciences-Volume*, Chicago, IL, Vol. 1, 842–848.

Tafreshi, D., and Racine, T. P. (2016). Children's interpretive theory of mind: the role of mothers' personal epistemologies and mother-child talk about interpretation. *Cogn. Dev.* 39, 57–70. doi: 10.1016/j.cogdev.2016.04.003

Weinstock, M. (2016). "Epistemic cognition in legal reasoning," in *Handbook of Epistemic Cognition*, eds J. A. Greene, W. A. Sandoval, and I. Bråten (New York, NY: Routledge), 215–229.

Weinstock, M. (2018). "The developmental roots of knowing: interpretive theory of mind and subjectivist epistemology," *Paper Presented at the Biennial Meeting of the European Association for Research on Learning and Instruction SIGs* 20 and 26, Jerusalem.

Weinstock, M., and Cronin, M. A. (2003). The everyday production of knowledge: individual differences in epistemological understanding and juror reasoning skill. *Appl. Cogn. Psychol.* 17, 161–181. doi: 10.1002/acp.860

Weinstock, M. P., Neuman, Y., and Glassner, A. (2006). Identification of informal reasoning fallacies as a function of epistemological level, grade level, and cognitive ability. *J. Educ. Psychol.* 98, 327–341. doi: 10.1037/0022-0663.89.2.327

Wellman, H. M., Fang, F., and Peterson, C. C. (2011). Sequential progressions in a theory-of-mind scale: longitudinal perspectives. *Child Dev.* 82, 780–792. doi: 10.1111/j.1467-8624.2011.01583.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Weinstock, Israel, Cohen, Tabak and Harari. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 14 July 2020
doi: 10.3389/fpsyg.2020.01737

The Impact of a Construction Play on 5- to 6-Year-Old Children's Reasoning About Stability

Anke Maria Weber, Timo Reuter and Miriam Leuchter*

Institute for Children and Youth Education, Educational Sciences, University of Koblenz and Landau, Landau, Germany

Theory: Young children have an understanding of basic science concepts such as stability, yet their theoretical assumptions are often not concerned with stability. The literature on theory theory and theory-evidence coordination suggests that children construct intuitive theories about their environment which can be adjusted in the face of counterevidence that cannot be assimilated into the prior theory. With increasing age, children acquire a Center theory when balancing objects and try to balance every object at their middle, succeeding with symmetrical objects. Later, they acquire the basic science concept of stability through learning that the weight distribution of an object is of importance. Thus, they acquire a Mass theory and succeed in balancing asymmetrical objects as well. Fluid and crystallized intelligence might contribute to children's acquisition of Mass theory. Moreover, their Mass theory might be supported by implementing a playful intervention including (a) material scaffolds and (b) verbal scaffolds.

Aims: We investigated which theories children have about stability and whether these theories can be adjusted to Mass theory by implementing a playful intervention.

Method: A total of 183 5- to 6-year-old children took part in the study with a pre-post-follow-up intervention design. Children's Mass theory was assessed with an interview in which children explained constructions' stabilities. The children received a playful intervention with two differing degrees of scaffolding (material scaffolds or material + verbal scaffolds) or no scaffolding.

Results: At first few children used a Mass theory to explain their reasoning. However, after being confronted with counterevidence for the asymmetrical constructions, children changed their explanation and applied a Mass theory. More children in the play group with the highest degree of scaffolding, i.e., material + verbal scaffolds, acquired a Mass theory compared to the other groups. Fluid as well as crystallized intelligence contributed to children's acquisition of a Mass theory.

Discussion: Counterevidence can support children in their acquisition of a Mass theory. A playful intervention with scaffolding supports children even more.

Keywords: guided play, theory theory, theory-evidence coordination, free play, science learning, intelligence

OPEN ACCESS

Edited by:
Christopher Osterhaus, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Horst Krist, University of Greifswald, Germany
Jennifer M. Zosh, Penn State University, Brandywine, United States

****Correspondence:***
Anke Maria Weber
weber-a@uni-landau.de

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *31 January 2020*
Accepted: *23 June 2020*
Published: *14 July 2020*

Citation:
Weber AM, Reuter T and Leuchter M (2020) The Impact of a Construction Play on 5- to 6-Year-Old Children's Reasoning About Stability. Front. Psychol. 11:1737. doi: 10.3389/fpsyg.2020.01737

INTRODUCTION: SCIENCE EDUCATION

Early knowledge acquisition in the science domain and scientific reasoning lay the foundation for the understanding of complex science concepts, which are relevant throughout the school years and in later life (Eshach and Fried, 2005; Trundle and Saçkes, 2012). Accordingly, studies have demonstrated that science learning and scientific reasoning can be promoted in the early years of childhood (Gelman and Brenneman, 2004; Akerson et al., 2011; Klahr et al., 2011; Cremin et al., 2015).

Children construct intuitive theories to explain what is happening around them, and adjust these theories continuously (Gopnik and Wellman, 2012). These theories encompass science concepts, which might be altered by confronting children with counterevidence (Bonawitz et al., 2012). Promoting young children's science learning aims at helping them adjust their theories and should consider children's developmental constraints by considering children's everyday activities, e.g., their play (Copple and Bredenkamp, 2009). One possibility for such science-related play could be construction play in the form of block play, which is an important leisure activity for young children (Rubin et al., 1978; Pellegrini and Gustafson, 2005; Borriello and Liben, 2018; Verdine et al., 2019). An adult's guidance might be integrated into children's play in the form of scaffolding, which might support children's science learning (cf. van de Pol et al., 2010; Klahr et al., 2011; Mantzicopoulos et al., 2013).

Therefore, children's science theories might be supported through science-related play that focuses on children's experiences and encompasses an adult's scaffolding.

Theory Theory, Bayesian Inference, and Theory-Evidence Coordination

Fostering scientific reasoning is one goal of science education (Chin, 2007; Klahr et al., 2011). Studies on children's scientific reasoning rely on literature concerned with theory theory, Bayesian inference and theory-evidence coordination. Theory theory is concerned with the adjustment of children's intuitive theories when they are confronted with evidence. Bayesian inference focuses on the interaction of intuitive theories with evidence, while theory-evidence coordination investigates the conditions under which children can interpret evidence.

According to theory theory, children construct intuitive theories about their environment, which have similarities with scientific theories (Gopnik and Wellman, 2012). Intuitive and scientific theories share at least five characteristics. (1) They encompass causal representations of the surrounding world, (2) may be hierarchically organized, (3) provide possible explanations for regularities, (4) allow predictions of regularities, and (5) can be adjusted in the face of counterevidence. In this process, not only the explanations for certain subordinate relations but also the general assumptions about regularities can change (Bonawitz et al., 2012; Gopnik and Wellman, 2012). Such adjustment occurs, often gradually, if a child is confronted with counterevidence that cannot be assimilated into their prior theory, either through their own interventions, e.g., in their play, or through observing the interventions of others (Gopnik and Wellman, 2012). According to Gopnik and Meltzoff (1997), children pass through the same developmental processes and therefore have similar representations about the same objects at roughly similar times in their lives. Most children do not test their theories with experiments but adjust them when they are confronted with evidence (Gopnik, 2013). Research has provided support for theory theory, *inter alia*, in the domain of balance (Bonawitz et al., 2012), and biology (Schulz et al., 2007).

Researchers have applied Bayesian inference to the theory theory framework to value the role of probability and prior knowledge on learning processes (Schulz et al., 2007; Gopnik and Wellman, 2012; Gopnik, 2013). Bayesian inference indicates how a learner changes their theory after being confronted with a set of evidence and how children might combine theory and evidence (Schulz et al., 2007). Bayesian models consider prior knowledge and its effect on inductive reasoning as well as how much a person believes one theory to be true. Furthermore, prior beliefs and evidence might interact, e.g., a child might interpret data according to their prior beliefs and dismiss counterevidence (Hawkins et al., 1984; Griffiths et al., 2011; Gopnik, 2013). Children with consistent presumptions will likely change their theory less easily than children with inconsistent presumptions (Gopnik et al., 2001; Sobel et al., 2004; Griffiths et al., 2011).

Studies on theory-evidence coordination have found that young children often face problems relating a theory with evidence (Ruffman et al., 1993; Koerber et al., 2005; Piekny and Maehler, 2013), which seemingly contradicts the results of studies on theory theory (Schulz et al., 2007; Bonawitz et al., 2012). However, taking a closer look at the studies on theory-evidence coordination, these studies showed that young children primarily face problems when evidence was presented in the form of imperfect covariation. Children were more likely to successfully coordinate theory with evidence when the evidence was presented in the form of perfect covariation, which is how the studies on theory theory presented evidence. For example, Sobel et al. (2004) found that children were even able to make inferences from indirect evidence of perfect covariation in the form of data they had not directly observed.

In conclusion, at least three factors might contribute to young children's science learning with regard to their developmental constraints and should be considered. (1) Children can interpret perfect covariation but face problems with imperfect covariation (Koerber et al., 2005). Therefore, evidence should be presented in the form of perfect covariation. (2) Children have prior theories about science phenomena and often have similar theories at a certain age (Gopnik and Meltzoff, 1997). Therefore, these prior theories should be considered so that the presented evidence relates to children's intuitive theories. (3) Children with consistent prior theories might need to see counterevidence repeatedly to adjust their theory because this adjustment often happens gradually (Gopnik and Wellman, 2012). Therefore, children should receive enough time to deal with the phenomenon.

The question remains how best to confront young children with evidence relating to their science theories. Children's developmental constraints can be addressed by allowing for

activities that occur in their everyday lives, e.g., play (Copple and Bredenkamp, 2009). Moreover, play might be enriched by scaffolding materials as well as an adult's verbal support (Zosh et al., 2018).

Material and Verbal Scaffolds in Guided Play

Play-based learning is the mandated pedagogy in early years' curricula in many countries (Hirsh-Pasek et al., 2009; Pyle et al., 2017) and is regarded as developmentally appropriate practice (Copple and Bredenkamp, 2009). Researchers have widely agreed that play can be characterized as a voluntary, intrinsically motivated, child-directed, and process- rather than goal-oriented behavior that contains elements of choice (Rubin et al., 1983; Pellegrini, 2013; Weisberg et al., 2013; Trawick-Smith, 2015; Daubert et al., 2018).

Zosh et al. (2018) define play as a spectrum that allows for different *types of play*, ranging from *free play* as voluntary, intrinsically motivated, and process-oriented behavior directed by the child to more goal-oriented and adult-directed forms of *playful instruction*. In between these two poles, *guided play* represents a blend of *free play* and *playful instruction* (Zosh et al., 2018). Guided play can be described as a playful activity that is directed by the child, i.e., the child is autonomous to decide what to do, for how long and at what pace. The adult's role in guided play is to prepare a play environment and support the children's activities to facilitate learning (Weisberg et al., 2013, 2016; Zosh et al., 2018).

Guided play shares strong commonalities with the guided inquiry principle, which has been identified as one of the most effective approaches to learning and teaching (Mayer, 2004; Alfieri et al., 2011; Lazonder and Kamp, 2012). Researchers have frequently framed guidance in inquiry-based science teaching within the scaffolding construct (Hmelo-Silver et al., 2007). In the scaffolding literature, both material scaffolds and an adult's verbal scaffolds are considered effective in guiding children's learning (van de Pol et al., 2010; Martin et al., 2019). Accordingly, guided play can take at least two different forms (Weisberg et al., 2016), guided play with *material scaffolds only*, and guided play with *additional verbal scaffolds*.

In guided play with *material scaffolds*, the adult provides the children with purposefully designed and structured materials aiming at a specific learning objective (Weisberg et al., 2016). Research indicates that children's explorations with purposefully structured and limited materials can foster science learning (Cook et al., 2011; van Schijndel et al., 2015). In particular, learning materials are effective when they link the learning objective to children's prior knowledge (Leuchter et al., 2014) and focus children's attention on those aspects that are essential for understanding (DeLoache, 2014). For example, to foster children's stability concepts, the adult might provide the children with an assembly of building blocks and a variety of photographs. In guided play with material scaffolds only, the adult initiates the play activities by inviting the children to rebuild the block constructions depicted on the photographs and to explore whether the constructions remain stable or tumble. By building these constructions, the children are likely to face evidence (the construction remains stable or tumbles) that might be incompatible with their intuitive theory (cf. Gopnik and Wellman, 2012). However, beyond initiating children's explorations, the adult does not intervene in the process.

Research suggests that children show more explorative behaviors when an adult takes a passive role (Bonawitz et al., 2011). In contrast, studies indicate that children's unguided explorations might not be sufficient to encounter the learning objective (Butts et al., 1994; Chen and Klahr, 1999; Klahr and Nigam, 2004; Sarama and Clements, 2009). In the stability example, children might rebuild the construction inappropriately and thus might witness incorrect evidence or imperfect covariation. Moreover, children might interpret evidence inappropriately to confirm their intuitive but incorrect theory.

In guided play with *additional verbal scaffolds*, the adult not only provides materials but additionally plays along with the children, supports the children's play verbally and encourages higher order thinking (Chin, 2007; Haden, 2010; Kleickmann et al., 2016; Weisberg et al., 2016; Martin et al., 2019). The adult can use a set of verbal scaffolding techniques to aid children's cognitive activities (for an overview cf. van de Pol et al., 2010) and support the cognitive processes involved in theory-evidence coordination, thus helping children adjust their intuitive theory. *Activating prior knowledge* by asking questions and prompting the children to express their presumptions, e.g., whether a block construction will remain stable or tumble can facilitate the integration of new aspects into existing schemata (Mayer, 1997; Weinert and Helmke, 1998; Gurlitt and Renkl, 2010). Additionally, *asking for the child's reasoning,* e.g., by prompting the child to justify their presumptions about the block construction's stability allows the child to structure their prior knowledge and thinking processes (Hsin and Wu, 2011). *Providing explanations* may help the child coordinate their observations with an evidence-based interpretation of a phenomenon (Murphy and Messer, 2000; Renkl, 2002). *Encouraging comparisons* supports the child in identifying relational similarities or differences between entities by highlighting certain features (Hsin and Wu, 2011; Richey and Nokes-Malach, 2013). Furthermore, *modeling*, i.e., performing certain behaviors and thinking styles, offers a possibility for imitation (Mayer, 2004; Hmelo-Silver et al., 2007).

Research indicates that guided play with *additional verbal scaffolds* promotes children's science learning more effectively than free play (Pine et al., 1999; Hadzigeorgiou, 2002; Fisher et al., 2011; Reuter and Leuchter, 2020). However, there are only a few studies that have deliberately compared the effectiveness of *material scaffolds* with *additional verbal scaffolds* for children's science learning. Leuchter and Naber (2019) found that a combination of structured materials and verbal scaffolds supported 6- to 7-year-old children's learning in the physics domain of force more than only materials, only verbal support or free exploration. Similarly, the results of Hadzigeorgiou (2002) show that 4- to 6-year-olds perform more meaningful activities at an inclined plane to explore the concept of mechanical stability when they received structured materials and verbal

scaffolding compared to children who received only materials or played freely.

Studying children's scientific reasoning in a playful context can aim at unraveling the interplay of material and verbal scaffolds. Concerning children's reasoning about science phenomena guided play can serve as a developmentally appropriate context to shed a light on (1) children's theory adjustment, (2) the way their prior theories interact with the evidence provided through the scaffolding materials, and (3) the conditions that may support children to coordinate theory with evidence.

The Statics Domain and Children's Beliefs About Balance

Statics can be defined as the state of equilibrium of an object, which in turn is concerned with forces acting on objects that are either at rest or in motion (Riley and Sturges, 1993). Statics is therefore concerned with stability. If the middle of a symmetrical object is supported by a supporting surface, the object will remain stable. Therefore, the consideration of an object's geometrical center is sufficient when rating symmetrical objects. For an asymmetrical object, however, the mass must be considered because the geometrical center and the center of gravity do not correspond. If the center of gravity of an object is supported, the object will remain in place; however, if it is not supported, the object will tumble. According to Bonawitz et al. (2012); Krist (2010), and Siegler (1976), with increasing age children develop an understanding of the weight distribution so that they can estimate the stability of an asymmetrical object/construction.

Studies with infants have mostly employed the violation of expectation paradigm and have shown that infants have basic knowledge about stability (Baillargeon and Hanko-Summers, 1990). Studies with older children, however, have shown that even preschoolers face problems explaining why certain objects either remain stable or tumble. Krist and Krüger (2012) explain this discrepancy with different approaches of the violation of expectation paradigm and verbal explanations as a possible reason for this ability gap. They state that being surprised (violation of expectation) does not take as much cognitive reasoning as verbally explaining one's underlying theory.

Young children indeed hold misconceptions about balance. Siegler (1976) and Siegler and Chen (1998) placed different weights at different distances on a fulcrum and asked children to rate the fulcrum's balance. The researchers found that children from 9 years of age started to consider both weight and distance, while younger children tend to view weight and distance separately. Other studies by Krist et al., have shown that between the ages of three to eight, children's abilities of balancing symmetrical and asymmetrical blocks and estimating symmetrical as well as asymmetrical objects' stabilities increase continuously independent of the type of assessment (rating photographs, Krist, 2010; eye tracking, Krist et al., 2018; balancing blocks, Krist et al., 2005). Even though children's estimation of asymmetrical blocks' stabilities increased, all three studies found children's performance on the estimation of symmetrical objects (e.g., cuboids) to be superior to their estimation of asymmetrical objects (e.g., L-shaped objects). As noted earlier, the center of gravity does not correspond to the geometrical center of an asymmetrical object. For symmetrical objects, however, considering their geometrical center is sufficient. Thus, children's difficulty in estimating the stability of asymmetrical objects indicates that they face problems considering the weight distribution.

Some studies have taken a closer look at children's theories about balance. Pine et al. (2007) asked 6- to 8-year-old children about their reasoning when balancing beams on a fulcrum and categorized their verbal utterances as well as their gestures into four categories: middle, weight, distance, and other. They found that most answers fell into the other or weight categories, and few children considered the distance. Moreover, Weber and Leuchter (2020) found that more than half of the 5- and 6-year-olds in their sample used an undifferentiated pattern when rating photographs of asymmetrical objects, approximately 1/3 applied Center knowledge, and less than 10% of children applied Mass knowledge.

The above studies have examined children's knowledge about stability from a developmental psychological perspective. However, it is also of interest if children's Mass knowledge can be supported in regard to their developmental constraints. Playful interventions with building blocks have supported the acquisition of different mathematical and spatial skills in other studies (e.g., Ferrara et al., 2011; Fisher et al., 2013; Borriello and Liben, 2018; Thomson et al., 2018; Verdine et al., 2019).

Regarding children's rating of stabilities, Pine and Messer (2003) found that 5- to 6-year-olds were able to balance more symmetrical as well as more asymmetrical blocks after playing with the blocks compared to a pretest. In another study, children between 4 and 7 years of age first balanced symmetrical and asymmetrical objects on a beam scale (Bonawitz et al., 2012), and their balancing behavior was categorized into three categories (*No*, *Center*, and *Mass theory*). Furthermore, the results indicated that younger children tend to use an undifferentiated pattern (No theory) and do not consider the center or the mass. Second, after balancing objects on a beam, children either played with a mass-consistent or a center-consistent toy on their own and freely. Afterward, they again balanced an asymmetrical block. Children who had a Center theory before playing observed evidence that did not confirm their theory if they were in the mass condition. Many of these children adopted a Mass theory. Children who had a Mass theory before playing also observed evidence that did not confirm their theory if they were in the center condition. Most of these children did not alter their balancing behavior and instead explained away the evidence and remained Mass theorists. This outcome indicated that even a short presentation of counterevidence can support children's learning, but that their prior theories need to be considered.

The different effects of free play and guided play with material and material + verbal scaffolds on children's science learning in the domain of balance with regard to their prior theories have not yet been investigated. Furthermore, it remains unclear whether these adjustments remain stable over a longer period of time or if the children relapse into their prior intuitive theories.

Possible Relationship With Intelligence

Intelligence is one of the most important prerequisites for learning. The ability to solve or complete puzzles or patterns is considered an indicator of fluid intelligence (Cattell, 1987; Flynn and Blair, 2013). Two components of fluid intelligence are figural perception and figural reasoning as indicators of an individual's ability to perceive and mentally represent objects and abstract certain characteristics (Cattell, 1987; Weiß and Osterland, 2013). In the context of stability, figural perception and figural reasoning might contribute to children's Mass theory. To rate stabilities correctly, children must perceive, mentally represent and abstract the spatial features of the objects or constructions (Weber and Leuchter, 2020).

Language capacity is considered an indicator of crystallized intelligence and is one of the key indicators of mental ability in young children (Cattell, 1987; Flynn and Blair, 2013). Language capacity contributes to knowledge acquisition (Thorsen et al., 2014). Moreover, children with a higher language capacity might find it easier to articulate their reasoning and might profit more from verbal scaffolds.

The Present Research

The present study is concerned with the effects of three different types of construction play on children's science learning in the statics domain, specifically constructions' stabilities. We implemented two types of guided play (verbal + material scaffolds, material scaffolds) and free play. Following the literature on theory theory, Bayesian inference, and theory-evidence coordination, young children's science learning may be fostered by confronting children with evidence in the form of perfect covariation (Koerber et al., 2005; Gopnik and Wellman, 2012). Furthermore, children's prior theories, which they have acquired through their everyday activities, e.g., their play, should be considered so that the presented evidence relates to these theories, which can then help children interpret the evidence (Gopnik and Meltzoff, 1997; Gopnik and Wellman, 2012). For example, at the age of 5 to 6, children might explain and predict the stability of an object with a Center theory or have other theories (Bonawitz et al., 2012; Weber and Leuchter, 2020). Material scaffolds can be prepared in such a way that they show perfect covariation for Mass theory and contradict Center theory. Through verbal scaffolding, an adult can help the children connect the evidence presented through the material scaffolds with their prior (intuitive) theories. Thus, scaffolds may support children's theory adjustment from Other[1] or Center theory to Mass theory. Since theory adjustment often happens gradually (Gopnik and Wellman, 2012), children should receive enough time to explore stabilities. We designed playful interventions that consider these constraints and investigated the effects of the different kinds of play on children's theory adjustments in the statics domain. Moreover, we explored whether these adjustments remained stable over an extended period of time.

[1] By Other, we do not imply that children have no theory at all, but rather treat it as a rest category for children's answers that were neither concerned with the center nor the mass of constructions (cf. Bonawitz et al., 2012, p. 221).

Finally, interindividual prerequisites might be partly responsible for children's theory adjustment and interact with the type of playful intervention that the children received. Research on theory theory, Bayesian inference and theory-evidence coordination suggests that children with a consistent prior theory might not adjust their theories as easily as children with an inconsistent prior theory (Koerber et al., 2005; Gopnik and Wellman, 2012). Thus, we are interested in the contribution of children's prior theories on their adjustments after being confronted with perfect evidence for Mass theory. Additionally, intelligence affects learning (Cattell, 1987; Flynn and Blair, 2013; Thorsen et al., 2014) and may thus contribute to theory adjustment as well. With respect to fluid intelligence, we hypothesize that figural perception and figural reasoning facilitate theory adjustment. With respect to crystallized intelligence, we hypothesize that children with higher language capacity might profit more from verbal scaffolds than children with lower language capacity.

Therefore, we specify the following research questions:

(1) Do children explain their reasoning about stability with Mass theory, Center theory or Other?
(2) Can guided play with material scaffolds and with or without verbal scaffolds enhance children's consistent use of Mass theory compared to free play (a) directly after a playful intervention and (b) over an extended period of time? Does intelligence relate to the consistent use of Mass theory?
(3) Does the consistency of children's prior theories relate to children's consistent use of Mass theory in the different play conditions (a) directly after a playful intervention and (b) over an extended period of time? Does intelligence relate to the consistent use of Mass theory when prior theories are considered?
(4) Do children with a consistent Mass theory after the playful intervention perform differently on a transfer test than children who did not use Mass theory consistently? Does intelligence relate to performance on the transfer test?

MATERIALS AND METHODS

Participants

In total, 183 children (88 girls, 95 boys), between the ages of five and six (M = 5.55, SD = 0.51), took part in the study. The participants visited 23 kindergartens in Germany (2 to 13 children per kindergarten), which all agreed to take part in the study and helped connect with the children and their parents. The kindergartens were located either in villages (700 to 3,000 inhabitants; N = 83 children), small cities (less than 20,000 inhabitants; N = 10 children) or medium sized cities (approximately 50,000 inhabitants; N = 91 children of whom 51 lived in the city center and 40 in the periphery). A total of 171 children were European, 9 were Asian, 2 were African, and 1 was Central American. All children and their parents were informed about the goals of the study, and all children took part voluntarily and with their parents' consent.

Some children dropped out of the study completely because, e.g., they moved or were ill on the dates agreed with the kindergartens, and some other children had missing values on some of the items. We used pairwise deletion because we decided to include the highest amount of data whenever possible. Therefore, the number of participants varies between different analyses.

Procedure

The study consisted of a pre-post-follow-up design with two guided play groups and a free play group. The pretest (T1) took place approximately 2 weeks before the play session and the immediate posttest (T2). The follow-up (T3) took place approximately 10 weeks after the posttest. The duration of the play session was approximately 1 h.

For the intervention, the children were parallelized into the three intervention groups according to their language capacity, which was assessed at pretest. Thus, matched samples were produced and each child in the Verbal group had a "language capacity twin" in the Material and in the Free play group. Both the *Material group* (59 children, 32 girls, and 27 boys) and the *Verbal group* (64 children, 27 girls, and 37 boys) received scaffolding materials in the form of building blocks. The Verbal group received additional verbal scaffolds. The *Free play group* (60 children, 29 girls, and 31 boys) played with building blocks freely. The reason for the differences in group size is that each intervention was to be conducted in a group of approximately four to six children to achieve ecological validity. Therefore, five children in the Verbal group were not assigned to language capacity triplets in the other two groups. In total, there were 51 interventions with group sizes varying between 2 to 6 children per group.

The play was led by one of six female experimenters. To prevent experimenter effects, group*experimenter was varied, so that all experimenters had led all intervention groups, i.e., Verbal, Material, and Free play group. In all intervention groups, the children were free to choose what they wanted to build or if they wanted to build with a friend or rather on their own. Furthermore, breaks were always possible, and the children were free to stop playing entirely (cf. Rubin et al., 1983; Weisberg et al., 2016). For manipulation check, the play sessions were video or audio recorded with the permission of parents and children. Based on the recordings, we rated children's playfulness according to Bundy et al. (2001) as well as children's on-task behavior. High inferential ratings showed that all children in all recordings showed indications of playfulness, e.g., children sang, laughed, joked around with another or the experimenter, chose to build challenging constructions, and built together. Moreover, children's on-task behavior did not differ between the groups.

Children in the Material group and the Verbal group received photographs of different block constructions, which differed in the number of blocks used and in their complexity. With each photograph came a box containing the building blocks needed for building the construction. The blocks differed in their shapes (cuboids, triangles, etc.), sizes, and colors (brown, black, yellow, red, and green). The materials were developed prior to this study and tested in play sessions to ensure that children were able to rebuild the structures shown on the photographs and had fun playing with the materials. The material scaffolds were implemented to structure the play and served as suggestions for the children. However, the children in the guided play groups were free to build constructions other than those we presented to them. In order for the activity to be enjoyable and playful for the children, we allowed the children to decide whenever they wanted to move on to the next activity. However, the experimenters could suggest another activity if they felt that the children started to lose interest.

The material scaffolds encompassed five activities and were presented in a standardized order (cf. **Supplementary Material 1** for all photographs):

(1) Black block (11 photographs): You can build the building shown on the photograph. Build the building and guess if the blocks remain stable or tumble.
(2) Add-a-block (8 photographs): These blocks on the photos were bewitched so they would remain stable. Can you rebuild the building, so that it is stable? (If a child did not succeed, the experimenter provided a green block): Look, here is a green block. Try to stabilize the building with it.
(3) Sliding (9 photographs): This is the sliding play. You rebuild the building on the photograph. Then, you slide the upper block along the lower block until it falls (experimenter models it). That's noisy, isn't it?
(4) Rebuild (11 photographs): You can just rebuild the building on these photographs and see how well you are doing. Some buildings are very easy to rebuild; others are more difficult. However, every single one will remain stable if built correctly.
(5) Stable/Tumble (8 photographs): The buildings on the photographs will sometimes remain stable, but at other times, the blocks are bewitched. Look at the photograph and predict "stable" or "tumble," and then try them out to see whether you were correct.

Additionally, the Verbal group received verbal scaffolds in the form of the activation of prior knowledge, asking for reasoning, the provision of explanations, the encouragement of comparisons, and modeling (cf. **Table 1**; Hogan and Pressley, 1997; van de Pol et al., 2010; for the script cf. **Supplementary Material 2**). The experimenters used this limited set of scaffolds presented in the script but applied them flexibly when playing with the children. All experimenters had received a training on how to apply scaffolding during play prior to leading the interventions.

The Free play group received a large box with the same building blocks as the guided play groups, but the blocks were unstructured. The experimenters did not suggest any buildings that the children could build but only told the children to play with the blocks freely.

During the play time, the experimenter praised the children's efforts in all three groups and motivated them to try again if they encountered problems with building. Sometimes children would ask the experimenter for help with building, which she would provide in the Verbal group. However, in the Material group or

the Free play group, she would friendly decline and suggest that children ask another child for help (cf. **Supplementary Material 3** for excerpts from the playful intervention).

Measures

Children's Theories

Children's theories about stability were assessed with a standardized single interview consisting of photographs of four symmetrical and four asymmetrical block constructions, which were always supported by a black block (cf. **Supplementary Material 4**). The children were asked to estimate whether the block construction presented in the first photograph would remain stable or not if the black block was removed. Afterward, the children received a total of five blocks, namely, four cuboid blocks consisting of two brown blocks, one black block and one yellow block (9 cm*3 cm*1.5 cm), as well as one smaller black cuboid block (3 cm*3 cm*1.5 cm). All the blocks were made of non-lacquered wood and colored by the researchers using acrylic colors to avoid slipperiness. The brown blocks had a narrow line drawn onto them with a pencil to facilitate finding the blocks' middle for the children. The children were asked to rebuild the construction presented in the photograph. Then, the interviewer repeated the children's former answer regarding the construction's stability (stable or unstable) and asked them to explain the prediction. They answered, and the interviewer invited them to remove the black block and ascertain whether they had rated the stability correctly. Then, they proceeded to the second block structure and so on. The interviews were videotaped or recorded if both the parents and the child had consented to it; if not, the interviewer made notes on the child's theory. The same interview was administered at each point of measurement, i.e., the same test items were presented at T1, T2, and T3. The testing time was approximately 10 min.

Only the asymmetrical block constructions were used in the data analyses. The test started with two symmetrical items to familiarize the children with the test logic. The fourth and sixth items showed symmetrical constructions and were applied to ensure that children had positive mastery experiences during the testing because studies have found that symmetrical items' stabilities are easier for children to estimate than asymmetrical items' stabilities (Krist, 2010). The three asymmetrical items showed perfect covariance for Mass theory because the weight distribution always determined the stability. However, the second asymmetrical item as well as all the symmetrical items could also be rated correctly with Center theory, while the first and third asymmetrical items could not. Therefore, the evidence for Center theory was imperfect.

The children's answers were coded following the speech coding scheme from Pine et al. (2007), as shown in **Table 2**. If a child was unable to verbalize their answer, but, e.g., pointed at the vertical block, their answer was also rated as Mass theory. If a child indicated the middle with gestures, the answer was rated as Center theory. The children's explanations were coded by two independent raters, Cohen's $\kappa > 0.90$.

Regarding the items that were not used for further analysis, the distribution of the children's answers at T1 was as follows. For the first item (familiarization, symmetrical), 18 Mass, 118 Center, and 39 Other. For the second item (familiarization, symmetrical), 14 Mass, 124 Center, and 37 Other. For the fourth item (motivation, symmetrical), 19 Mass, 115 Center, and 42 Other. For the sixth item (motivation, symmetrical), 15 Mass, 96 Center, and 64 Other. The seventh item showed an asymmetrical construction but was removed from further analyses because it was inconclusive. The probability of the item remaining stable was 50% for statics reasons. Thus, there is no definite answer to this item. The third, fifth, and eighth items were asymmetrical items that were included in the data analyses and used for the assessment of children's stability theories (cf. **Figure 1**).

Transfer Test

At the third point of measurement, a paper-pencil transfer test consisting of photographs of 16 asymmetrical block constructions was administered (cf. **Figure 2** and **Supplementary Material 5** for all transfer items), i.e., 8 stable constructions and 8 unstable constructions. The test was conducted in a group of up to six children who were seated back-to-back to

TABLE 1 | Scaffolding techniques used in the Verbal group (Hogan and Pressley, 1997; van de Pol et al., 2010).

Technique	Example
Activating prior knowledge	Have you ever seen something like this?
Asking for reasons	Can you explain this in more detail, so I can really understand what you think?
Providing explanations	Well done! If the heavy side of a block hovers in midair, the block will tumble
Encouraging comparisons	Your building looks different than [another child's building], doesn't it? What is different? Is something similar?
Modeling	Look! (Experimenter also looks very closely/experimenter shows how to build a certain building)

TABLE 2 | Coding scheme.

Coding	Speech	Example
Mass theory	The child refers to the weight being on one side of the brown blocks, mentions heaviness or talks about the importance of the vertical block	"This side is heavier." "It's because of the block that's standing on the other"
Center theory	The child refers to the middle of the block or a bigger amount of the block resting on either the black or the yellow block	"The brown block is resting more on the yellow block"
Other	Child speaks of something other than the two variables of interest (weight, middle), e.g., refers to the color	"I don't know." "It tumbles, because it tumbles"

FIGURE 1 | The asymmetrical constructions used to assess children's reasoning. Item 1 and 2 are stable constructions, item 3 is an unstable construction.

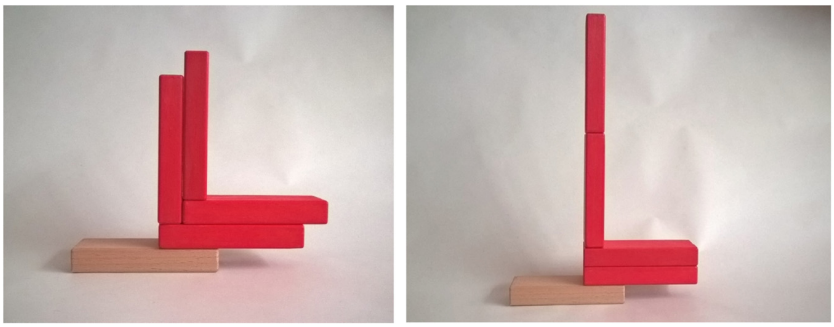


FIGURE 2 | Example items of the transfer test, both items show stable constructions.

prevent them from copying from each other. The test took approximately 10 min. The children were asked to rate the constructions' stabilities by either circling the photograph for a stable construction or crossing out the photograph for an unstable construction. Thus, children were not required to verbalize their knowledge. The constructions could only be rated correctly by considering the weight distribution because if the center of gravity was supported by the brown block but the geometrical center was not, then the constructions would always remain stable. However, if the center of gravity was not supported but the geometrical center was, then the construction would always tumble. For this instrument, the children did not need to explain their reasoning and only had to rate photographs. The children's content knowledge could be assessed to support the results of the reasoning test.

Aspects of Intelligence

Visual perception and figural reasoning as two aspects of fluid intelligence were assessed with the labyrinths and matrices subtests of the Culture Fair Test (CFT 1-R; Weiß and Osterland, 2013) at T1. *T*-values were not available, as only two subtests were conducted (for more information on test parameters, please cf. CFT 1- R handbook).

Language capacity as an indicator for crystallized intelligence was assessed with the German version of the Peabody Picture Vocabulary Test (PPVT 4; Dunn et al., 2015) at T1. The PPVT is a picture-based standardized single interview for which *t*-values are available. The test consists of 19 sets with 12 items each consisting of four pictures per item. Five-year-old children start with set 4, and 6-year-old children start with set 5. For each item, the children receive a word and point at the corresponding picture. This procedure is continued until the child answers 8 out of 12 items in a set incorrectly (cf. PPVT 4 handbook for more information). The PPVT 4 was also administered to ensure that all children understood and spoke the German language sufficiently.

The total testing time was approximately 1 h at T1, 10 min at T2 and 20 min at T3. At T3, the follow-up test was administered before the transfer test. During the testing, breaks were possible whenever the child or the experimenter considered it necessary.

Data Analyses

The statistics program R, version 3.6.2 (R Core Team, 2019) was used for data analyses. We used the survival (Therneau, 2015) package for the specification of survival analyses and the survminer (Alboukadel et al., 2019) package for forest plots.

In the first step, we investigated the number of children who had a Mass theory, a Center theory or Other on each point of measurement. The children received feedback about the correctness of their stability rating because after rating the stability, they removed the black block and could ascertain if they had answered correctly. Therefore, they had the opportunity to learn during testing. Thus, their answers were not independent and could not be summarized but instead were treated as individual events as the assumption of local stochastic independence was violated. Therefore, we used methods of risk-event analysis to analyze the group differences in the application of Mass theory (cf. Singer and Willett, 2003).

RESULTS

Children's Use of Mass Theory

To address the first research question concerned with children's use of causal relations, especially Mass theory, when explaining asymmetrical objects' stabilities, we investigated the percentage

of children who applied mass for explaining each of the three asymmetrical constructions on the first point of measurement. The following results were obtained across all groups. For the first item, 11%[2] of children explained their reasoning with Mass theory, 41% with Center theory, and 48% provided another answer. For the second item, 20% of children explained their reasoning with Mass theory, 56% with Center theory, and 23% provided another answer. For the third item, 16% of children explained their reasoning with Mass theory, 48% with Center theory, and 35% provided another answer, cf. **Figure 3** for percentage shares of children's theories (cf. **Supplementary Material 6** for percentages of correct answers). To compare the probability of answering with a specific theory between items, we compared the proportions of children who had answered with a specific theory (either Mass, Center or Other) using z-tests of proportions. They revealed that children were not more likely to explain their reasoning with Mass for any of the three items, $z = 5.41$, $df = 2$, $p = 0.067$; however, the answer probabilities for Center theory, $z = 8.87$, $df = 2$, $p = 0.012$, and Other, $z = 23.90$, $df = 2$, $p < 0.001$, differed across the items.

The percentage of children applying each theory for each group at T1, T2 and T3 is presented in **Table 3**. The use of Mass theory increased at T2 and T3, especially in the Verbal group.

Effects of Guided Play

To address the second research question concerning whether different kinds of play can enhance children's consistent use of Mass theory directly after the intervention as well as over a longer period of time, we used methods of survival analyses. Survival analysis is used to analyze the expected duration of time until an event takes place. In our case, the event is the children's consistent use of Mass theory after the playful intervention.

Consistency at T2

First, we used the binomial distribution to find a cut-off that guarantees that the probability of children finding the correct answer through guessing was below 10%. This enabled us to find how many correct answers might be given through guessing with $\sigma = 1.64$, i.e., $p < 0.10$, and a binomial probability of 1/3. Thus, we are able to categorize the children into children who explained their reasoning with Mass theory either consistently or inconsistently directly after the intervention at T2. Children with 3 out of 3 correct answers were rated as answering consistently, $p = 0.037$. Seven children who had explained their reasoning with Mass theory consistently at T1 (3 out of 3 Mass answers) were excluded from these analyses.

We defined each item as a point in time; therefore, time = 1 refers to the first item of T2, time = 2 to the second item of T2, and time = 3 to the third item of T2. The event of answering with Mass theory 3 out of 3 times could only take place at time = 3 or not at all. The detailed results of the Kaplan-Meier analysis are presented in **Supplementary Material 7**. The survival rate implies the percentage of children who remain either Center theorists or Other, and thus are not applying Mass theory consistently at T2. Therefore, the rate of children who applied Mass theory consistently is 100% minus survival rate, e.g., 100%-77% for the Verbal group. The results indicated that 23% of the children in the Verbal group explained their reasoning with Mass consistently at T2 compared to 9% of children in the Material group and 6% of children in the Free play group.

To investigate the results of the Kaplan-Meier analysis with a stricter procedure and to include the contributions of metric predictors, i.e., fluid and crystallized intelligence, a Cox regression (cf. **Table 4**) was specified, likelihood ratio test = 12.87, $p = 0.012$. The Cox proportional hazard model assumes that the hazard curves for the groups should be proportional. This means that if child 1 is twice as likely to explain their reasoning with Mass theory than child 2 at an initial point in time, then at all later points in time, child 1 remains twice as likely to explain their reasoning with Mass theory consistently compared to child 2. In this particular Cox-regression, the event could only take place at time = 3. Therefore, the proportional hazard assumption was met.

The Cox regression showed group differences in the consistency of answering with mass between the Free play group and the Verbal group. However, there were no differences between the Material group and the Verbal group or the Free play group and the Material group. The Verbal group was almost four times as likely to explain their reasoning with Mass theory consistently compared to the Free play group, as indicated by the regression coefficient, and by a factor of HR = 3.70. Although neither the differences between the Verbal group and the Material group nor those between the Material group and the Free play group were statistically significant, descriptively, the Cox regression implied that the Verbal group had twice the chance of applying Mass theory consistently than the Material group, HR = 1.99 and that the Material group had approximately twice the chance than the Free play group, HR = 1.86. Crystallized intelligence contributed to the consistent application of Mass theory, while fluid intelligence did not.

We tested whether crystallized intelligence interacted with the children's Mass theory in the three intervention groups. The analysis showed a difference between the Material group and the Verbal group dependent on crystallized intelligence, crystallized intelligence* ΔMaterial–Verbal, $b = 0.15$, $p = 0.032$. This indicates that children with high crystallized intelligence in the Verbal group profited more than children with high crystallized intelligence in the Material group (**Table 5**).

Consistency Over T2 and T3

Next, we were interested in whether children's answers differed in their consistency over a longer period of time to check if the effect of the guided play was lasting. Therefore, we combined the three items of T2 and T3 into 6 points in time. Again, we used the binomial distribution with $\sigma = 1.64$, $p < 0.10$, and binomial probability = 1/3 to categorize the children into children who explained their reasoning with Mass theory either consistently or inconsistently. Children with ≥4 Mass explanations out of 6 when combining the items of the posttest and the follow-up were categorized as answering consistently, $p = 0.097$. The first point in time on which the event could take place was time = 4, i.e., the first item of T3, because the children had to answer four items with Mass theory to fulfill the event. The

[2]Rounded percentages.

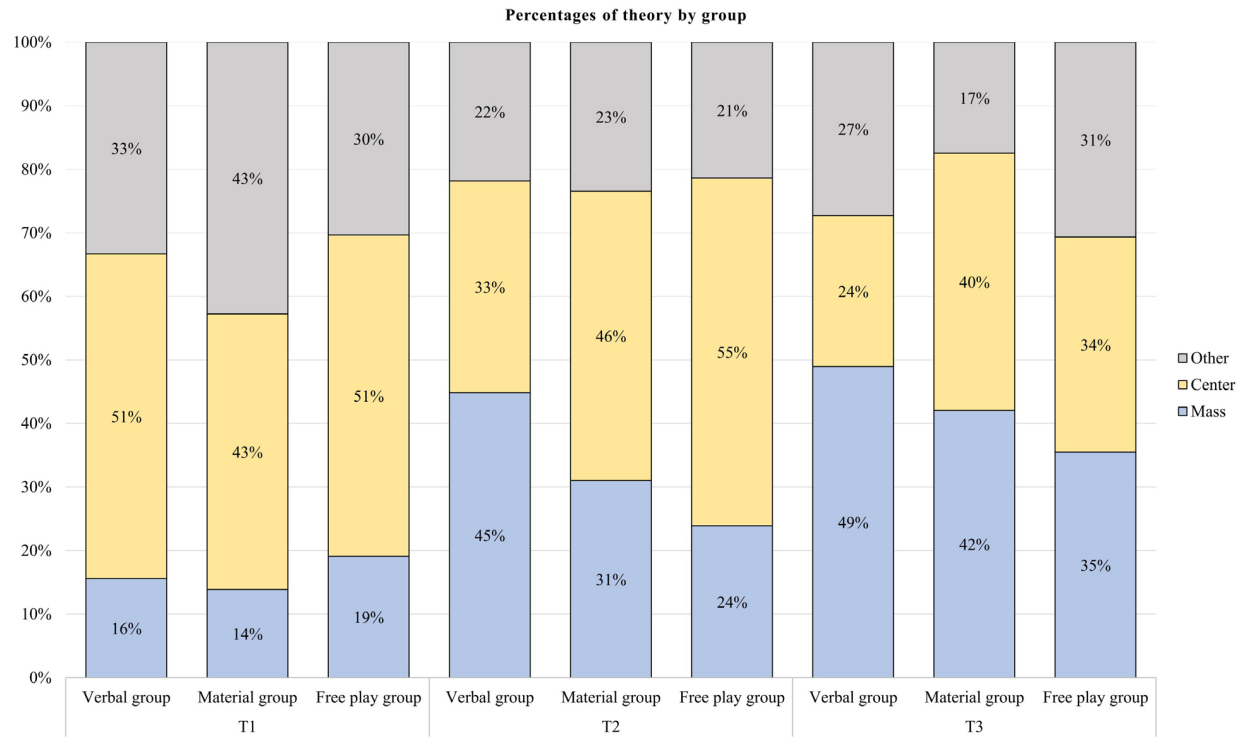


FIGURE 3 | Percentages of theories applied by the children to explain all three reasoning items' stabilities per measurement point and intervention group.

TABLE 3 | Number of children applying each theory in each group over all three points of measurement.

		Verbal group (*N* = 64)			Material group (*N* = 59)			Free play group (*N* = 60)		
	Item	**Other (%)**	**Center (%)**	**Mass (%)**	**Other (%)**	**Center (%)**	**Mass (%)**	**Other (%)**	**Center (%)**	**Mass (%)**
T1	1	47	40	13	51	40	9	47	42	12
	2	24	60	16	30	50	20	15	59	25
	3	29	53	18	49	40	11	29	51	20
T2	1	38	25	38	32	43	25	34	51	15
	2	14	39	46	13	53	34	9	60	30
	3	13	36	51	25	40	35	21	53	26
T3	1	35	23	42	40	36	24	48	29	23
	2	15	27	58	10	36	55	8	46	46
	3	32	21	47	2	50	48	23	54	23

Percentages are given in rounded numbers.

event could also take place at time = 5, i.e., the second item of T3, and at time = 6, i.e., the third item of T3. We specified a Kaplan-Meier analysis to investigate the percentage of children using Mass theory consistently (cf. **Figure 4** and **Supplementary Material 7**). Extending the descriptive results, the children in the Verbal group had the highest percentage of using Mass theory consistently at each point in time compared to the other two groups, Verbal group = 40%, Material group = 23%, Free play group = 15%.

Next, a Cox regression (cf. **Table 5**) was specified, likelihood ratio test = 22.38, $p < 0.001$. First, we tested the proportional hazard assumption by correlating the scaled Schoenfeld residuals for group membership with time to ensure that the time and the residuals were independent. The hazard curves for the groups were proportional, as indicated by the global test, $\chi^2 = 5.56$, $p = 0.234$, as well as the group comparisons, all $p > 0.05$. The Cox regression showed group differences between the Verbal group and the Free play group, with the Verbal group having a higher chance of explaining their reasoning with Mass theory consistently by factor HR = 3.45. Again, there were no group differences between the Material group and the Free play group or between the Material group and the Verbal group. However, descriptively, the Verbal group had the highest chance of explaining their reasoning with Mass theory. Fluid and crystallized intelligence contributed to the consistent explanation with Mass theory. For the hazard ratios, cf. **Figure 5**.

TABLE 4 | Cox-regressions for children's acquisition of Mass theory at T2.

					95% CI	
	b	*SE*	*z*	HR	LL	UL
Development of consistencies between groups (Cox-regression)						
ΔFree play–Verbal	1.31*	0.65	2.01	3.70	1.03	13.26
ΔMaterial–Verbal	0.69	0.58	1.18	1.99	0.63	6.26
ΔFree play–Material	0.62	0.76	0.81	1.86	0.42	8.31
Fluid intelligence	0.04	0.06	0.69	1.04	0.93	1.16
CrI	0.08*	0.03	2.41	1.08	1.01	1.15
Interaction of crystallized intelligence with intervention group						
ΔFree play–Verbal*CrI	−0.05	0.11	−0.44	0.95	0.76	1.19
ΔMaterial–Verbal*CrI	0.15*	0.07	2.15	1.67	1.01	1.34
ΔFree play–Material*CrI	−0.20	0.12	−1.71	0.82	0.65	1.03
CrI free play	0.16	0.10	1.58	1.18	0.96	1.44
CrI material	−0.04	0.06	−0.68	0.96	0.86	1.08
CrI verbal	0.11**	0.04	2.65	1.12	1.03	1.22
Fluid intelligence	0.05	0.06	0.87	1.05	0.94	1.18

*CrI, crystallized intelligence; HR, hazard ratio; CI, 95% confidence-interval; LL, lower level; UL, upper level. *p < 0.05. **p < 0.01.*

TABLE 5 | Cox-regressions for children's acquisition of Mass theory at T3 considering T2.

					95% CI	
	b	*SE*	*z*	HR	LL	UL
Development of consistencies between groups (Cox-regression) over T2 and T3						
ΔFree play–Verbal	1.22**	0.45	2.74	3.40	1.42	8.16
ΔMaterial–Verbal	0.57	0.41	1.39	1.77	0.79	3.95
ΔFree play–Material	0.66	0.51	1.30	1.93	0.72	5.18
Fluid intelligence	0.10*	0.04	2.48	1.11	1.02	1.20
CrI	0.06*	0.02	2.56	1.06	1.01	1.11
Interaction of crystallized intelligence with intervention group						
ΔFree play–Verbal*CrI	−0.02	0.06	−0.37	0.98	0.87	1.10
ΔMaterial–Verbal*CrI	0.14*	0.05	2.57	1.15	1.03	1.28
ΔFree play–Material*CrI	−0.16*	0.07	−2.30	0.85	0.74	0.98
CrI free play	0.11*	0.05	1.99	1.11	1.00	1.23
CrI material	−0.06	0.05	−1.25	0.94	0.86	1.03
CrI verbal	0.08**	0.03	2.61	1.09	1.02	1.16
Fluid intelligence	0.11**	0.04	2.61	1.12	1.03	1.21

*CrI, crystallized intelligence; HR, hazard ratio; CI, 95% confidence-interval; LL, lower level; UL, upper level. *p < 0.05. **p < 0.01.*

Again, the interaction of crystallized intelligence and the intervention group was included in the Cox regression (cf. **Table 5**). Crystallized intelligence contributed to the consistent use of Mass theory in the Verbal group and in the Free play group but not in the Material group. These differences were statistically significant. For low crystallized intelligence, the children in the Material group profited most from the intervention compared to the Verbal group and the Free play group.

Relationship of Children's Theory at T1 and Children's Consistent Use of Mass Theory

To address the third research question concerned with the relationship of the children's prior theories and their consistent use of Mass theory after the playful intervention, we categorized the children into those answering consistently or inconsistently at T1. For this method, we used the same criterion used for the prior analyses, i.e., children explaining their reasoning with either Center theory or Other 3 out of 3 times, $\sigma = 1.64$, $p < 0.10$, binomial probability of 1/3, were categorized as answering consistently at T1, and the other children were categorized as answering inconsistently. Hence, for the following analyses, the sample was divided into six groups, i.e., a consistent and inconsistent group for each of the three intervention groups. For categorizing children consistently answering with Other, we considered those children who had provided a theory neither concerned with the center nor the mass of constructions for all three items of the pretest.

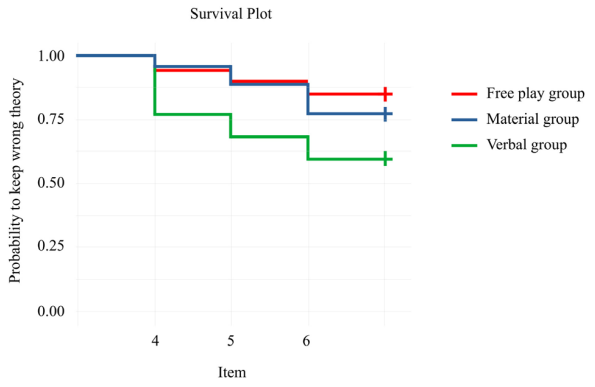


FIGURE 4 | Survival curves of the three play groups.

For the survival analysis, we applied the criterion of children explaining their reasoning with ≥4 Mass out of 6 to investigate whether children's prior theories relate to their acquisition of Mass theory. Kaplan-Meier analysis indicated that the children in the Verbal group who had answered inconsistently at T1 had the highest chance of explaining their reasoning with Mass consistently after the guided play, with 62% (cf. **Supplementary Material 7** for the comprehensive results).

Next, a Cox regression (cf. **Table 6**) was specified, likelihood ratio test = 26.15, $p < 0.001$. The hazard curves for the groups were proportional, as indicated by the global test, $\chi^2 = 7.13$, $p = 0.416$, as well as the group comparisons, all $p > 0.05$. We decided to use the Verbal group children who had answered inconsistently at T1 as the reference group for the Cox regression because theory suggests that this group should have the highest probability of explaining their reasoning with Mass theory. We found that this group had a significantly higher probability of explaining their reasoning with Mass theory than the Free play group children who had answered inconsistently at T1. Furthermore, descriptively, the Verbal group children who had answered inconsistently at T1 had the highest probability of all groups for explaining their reasoning with Mass theory (cf. **Figure 6**). We found no differences between the Free play group children who had answered consistently at T1 and any of the other groups. Fluid and crystallized intelligence contributed to the consistent use of Mass theory for all groups.

Transfer Test

To address the fourth research question concerned with children's performance on the transfer test, we compared children who had explained their reasoning with Mass theory consistently at T2 and T3, $M = 11.44$, $SD = 3.50$, to children who had not explained their reasoning with Mass theory consistently at T2 and T3, $M = 7.41$, $SD = 4.16$. Regardless of the intervention group, those children who had explained their reasoning with Mass theory consistently at T2 and T3 performed better on the transfer test than the other children, $t(65.83) = -5.26$, $p < 0.001$. A multiple regression analysis showed that neither fluid, $b = 0.05$, $p = 0.571$, nor crystallized intelligence, $b = 0.04$, $p = 0.319$, contributed to children's performance on the transfer test beyond the consistent use of Mass theory. In addition, crystallized intelligence did not moderate the consistent use of Mass theory on children's performance on the transfer test, $b = 0.17$, $p = 0.082$.

Furthermore, we compared the three intervention groups on the transfer test: Verbal group, $M = 8.17$, $SD = 4.64$; Material group, $M = 8.89$, $SD = 4.19$; Free play group, $M = 8.57$, $SD = 4.25$. Crystallized intelligence did not moderate the effect of the intervention group on the children's performance on the transfer test. An ANOVA showed no differences between the groups on performance in the transfer test, $F(2) = 0.27$, $p = 0.762$. Furthermore, there were no differences between the groups if the consistency at T1 was included, $F(5) = 1.37$, $p = 0.242$. Crystallized intelligence did not moderate the effect of consistency at T1*intervention group on performance in the transfer test.

DISCUSSION

Science learning in early childhood can and should be promoted (Eshach and Fried, 2005; Trundle and Saçkes, 2012). However, studies on early science learning are quite sparse, and it remains unclear how to best support young children with different individual prerequisites.

Therefore, we conducted a study on 5- to 6-year-olds' science learning in the specific domain of statics with regard to their prior intuitive theories and their individual cognitive prerequisites to investigate the effects of different types of play on theory adjustments. First, we were interested whether children explained their reasoning with Mass theory. In accordance with Pine et al. (2007), the children in our study faced problems estimating the stability of asymmetrical constructions and explaining the reasons for these stabilities. Most children provided another explanation and referred to characteristics of building blocks that have nothing to do with the mass or the center. Some children considered the center, and few considered the mass. This is in line with findings from other studies concerning the development of children's knowledge of mass (e.g., Siegler, 1976; Siegler and Chen, 1998; Krist, 2010; Bonawitz et al., 2012; Weber and Leuchter, 2020).

The result that young children do not have a Mass theory, but rather a Center theory or provide answers unconcerned with mass and center can serve as a starting point for designing learning environments that foster children's scientific reasoning, i.e., their theory adjustment, by providing them with perfect evidence for Mass theory (cf. Koerber et al., 2005; Klahr et al., 2011; Gopnik and Wellman, 2012). These learning environments should consider developmentally appropriate practice, i.e., play with scaffolds (Copple and Bredenkamp, 2009; Weisberg et al., 2016). Thus, we investigated whether a playful intervention could support children's theory adjustment from an intuitive prior theory to a Mass theory. Play was implemented in the form of construction play with building blocks with differing amounts of adult guidance. The Free play group played with blocks on their own. The Material group received static material scaffolds prepared by an adult. In both of these groups, the adult only motivated and praised the children's efforts but did not

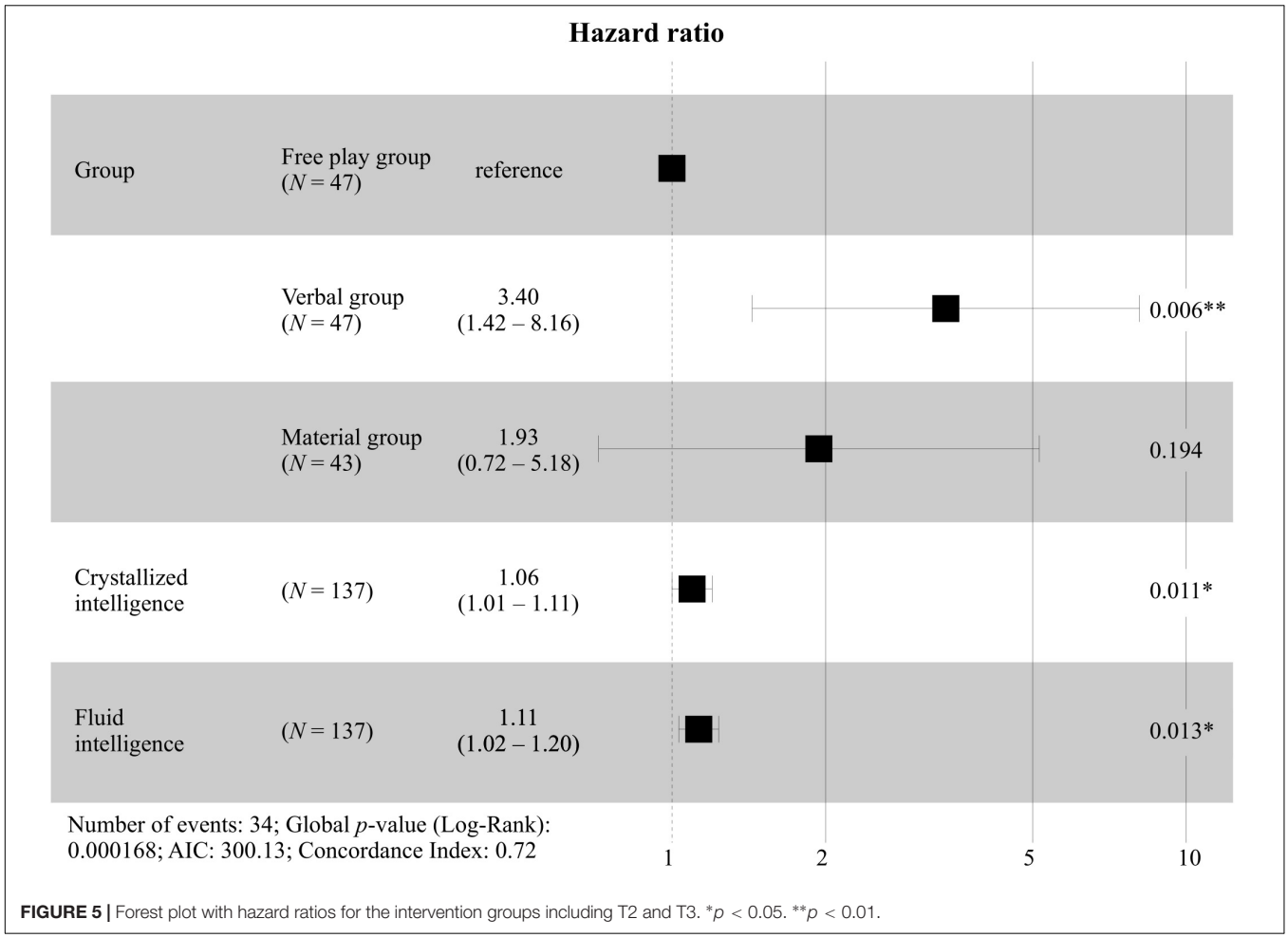


FIGURE 5 | Forest plot with hazard ratios for the intervention groups including T2 and T3. *$p < 0.05$. **$p < 0.01$.

TABLE 6 | Development of Mass theory consistencies between groups (Cox-regression) at T3 considering T2.

					95% CI	
	b	***SE***	***z***	**HR**	**LL**	**UL**
Verbal inconsistent as the reference group						
ΔVerbal inconsistent– Verbal consistent	−1.46	0.77	−1.90	0.23	0.05	1.04
ΔVerbal inconsistent– Material inconsistent	−0.66	0.47	−1.40	0.52	0.21	1.30
ΔVerbal inconsistent– Material consistent	−0.66	0.58	−1.15	0.52	0.17	1.60
ΔVerbal inconsistent– Free play inconsistent	−1.59**	0.53	−3.03	0.20	0.07	0.57
ΔVerbal inconsistent–Free play consistent	−1.11	0.76	−1.45	0.33	0.07	1.47
FI	0.08*	0.04	2.05	1.09	1.00	1.18
CrI	0.06**	0.02	2.67	1.06	1.02	1.10
Free play consistent as the reference group						
ΔFree play consistent–Verbal inconsistent	1.11	0.76	1.45	3.02	0.68	13.46
ΔFree play consistent–Verbal consistent	−0.35	1.01	−0.35	0.70	0.10	5.07
ΔFree play consistent–Material inconsistent	0.45	0.81	0.55	1.56	0.32	7.64
ΔFree play consistent–Material consistent	0.45	0.87	0.51	1.56	0.28	8.65
ΔFree play consistent–Free play inconsistent	−0.49	0.85	−0.58	0.61	0.12	3.22
FI	0.08*	0.04	2.05	1.09	1.00	1.18
CrI	0.06**	0.02	2.67	1.06	1.02	1.10

*FI, fluid intelligence; CrI, crystallized intelligence; Δβ, difference in regression coefficient between two groups. HR, hazard ratio; CI, 95% confidence-interval; LL, lower level; UL, upper level. *$p < 0.05$. **$p < 0.01$.*

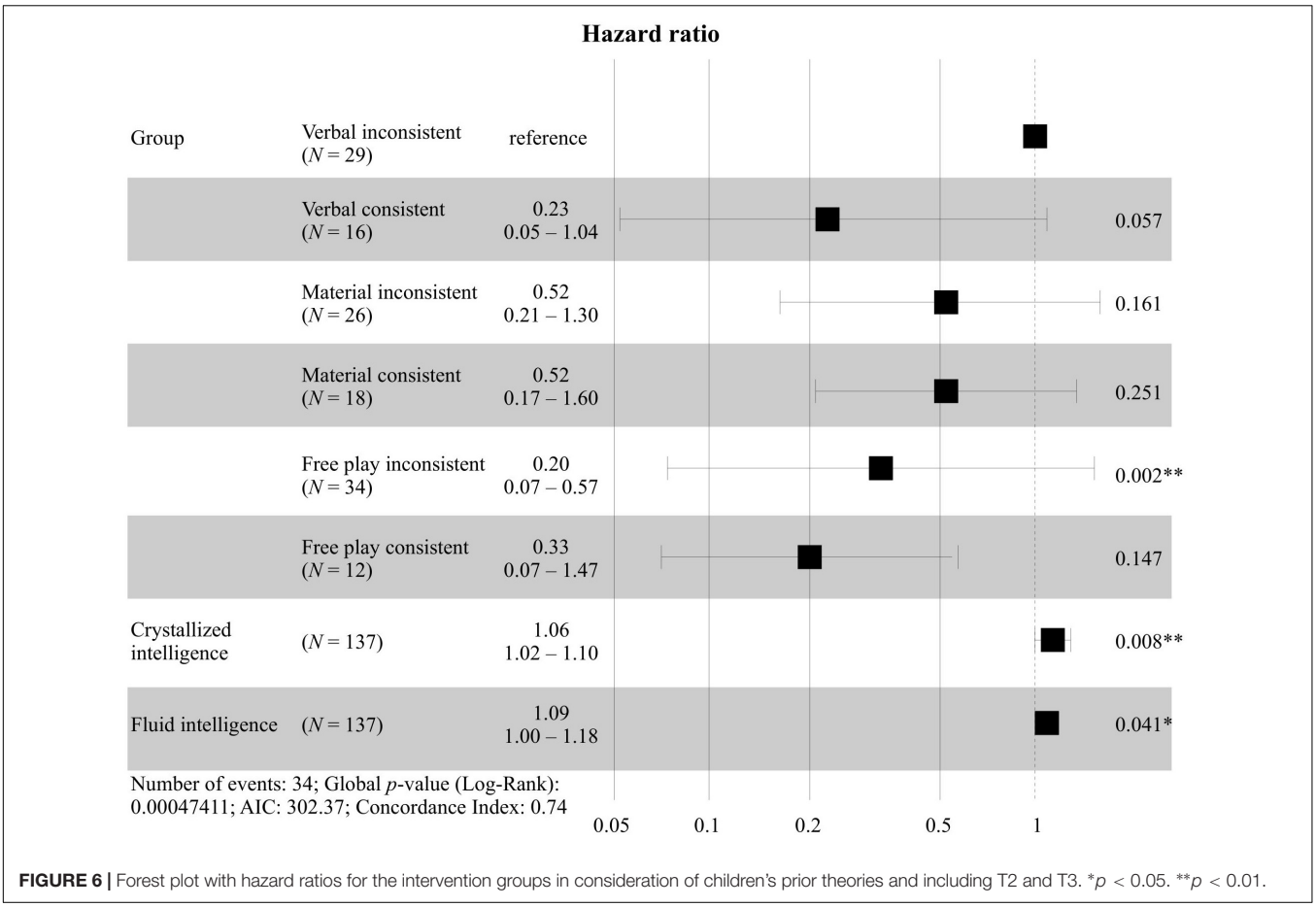


FIGURE 6 | Forest plot with hazard ratios for the intervention groups in consideration of children's prior theories and including T2 and T3. *$p < 0.05$. **$p < 0.01$.

intervene in the play process. The Verbal group received the same material scaffolds as the Material group, and additionally, an adult used verbal scaffolds during the play. Thus, the Verbal group received the highest amount of adult guidance. After the playful intervention as well as after the follow-up test, the Kaplan-Meier analysis showed that the children in the Verbal group were most likely to use Mass theory consistently to explain their reasoning. The Material group was more likely to use Mass theory than the Free play group. Group comparisons with a Cox regression showed that the Verbal group outperformed the Free play group but not the Material group.

The acquisition of Mass theory however, might be dependent on interindividual variables such as intelligence and consistency of prior theory that interact with the degree of scaffolding. Thus, we investigated whether children's prior theories are related to the adjustments of their theories to Mass theory. The link of children's prior theories with theory adjustment to Mass theory seems to be partly dependent on intervention type. The children in the Verbal group who had answered inconsistently during the pretest were most likely to adopt a Mass theory after the playful intervention compared to all other groups.

Last, we investigated whether the use of Mass theory to explain stabilities was related to children's results on a paper-pencil transfer test. Mass theorists performed better on the transfer test than Center theorists and children in the Other category. Our findings contribute to the literature on science education in the kindergarten years and will be discussed following the order of the research questions.

Children's Use of Mass Theory

The first research question was concerned with children's explanations of asymmetrical constructions' stabilities before the playful intervention. We investigated whether 5- to 6-year-old children explained asymmetrical constructions' stabilities with Mass theory, Center theory or Other.

Considering the assumptions of theory theory (Gopnik and Wellman, 2012), we found that children's theories about stability encompassed causal relations, as the theories provided explanations for the stability of the constructions and allowed for predictions of whether a construction would be stable or unstable depending on how it is supported. Thus, a child with a Center theory believes that the middle of a construction needs to be supported and that the middle is the cause for a construction's stability. A child with a Mass theory believes that a construction's weight needs to be supported and that the weight distribution is the cause for a construction's stability. A child providing another explanation might have other ideas concerning the causal relationship between support and stability, e.g., the color. Since most children could not explain asymmetrical constructions' stabilities correctly, we can assume that children do not have a

Mass theory (e.g., Siegler, 1976; Siegler and Chen, 1998; Krist, 2010; Bonawitz et al., 2012; Weber and Leuchter, 2020).

Nevertheless, the children were more likely to use Center theory and less likely to provide another explanation in the course of the pretest. The reason might be that they received visual confirmation of the Center theory for the symmetrical items. Hence, the children might have inferred and generalized Center theory as an explanation for constructions' stabilities (Bonawitz et al., 2012). Therefore, the children might have acquired a Center theory instead of remaining in the Other category or kept their Center theory instead of adopting Mass theory.

Even though all the children had the opportunity to learn about the mass even during the pretest because they received feedback concerning the constructions' stabilities, the probability that the children would explain their reasoning with Mass theory remained the same across all three items at pretest. This outcome indicates that the presented evidence at pretest might not have been sufficient to acquire Mass theory. Even though the children observed perfect evidence for Mass theory (cf. Koerber et al., 2005), a short presentation and asking for explanations about the stability in a one-to-one setting was not enough to introduce the children to Mass theory. Since the children seemed to be unable to acquire an understanding of the mass that easily, construction play with varying degrees of structuring seems to be an appropriate approach to investigate whether the children could acquire an understanding of the mass during a play.

Effects of Guided Play

The second research question was concerned with the playful interventions' effects on children's consistent use of Mass theory and the possible relationship with intelligence.

From the results it can be concluded that the more support the children received when confronted with evidence, the more likely they were to adjust their Other or Center theory and to explain their reasoning with Mass theory. This result indicates that children need support when learning about stabilities. Guided play with material and verbal scaffolds has been shown to support children's acquisition of Mass theory more than free play (cf. Zosh et al., 2018). In the Material group, the children might have overlooked the evidence, and in the Free play group, the children could only have observed it randomly.

Consistent with the literature on theory theory and theory-evidence coordination (e.g., Koerber et al., 2005; Gopnik and Wellman, 2012), our results indicate that three factors should be fulfilled when supporting young children's learning about science: (1) perfect covariation of the evidence, (2) considering the children's prior theories, and (3) leaving the children with enough time to explore the evidence. We approached these factors by (1) taking care to present children with perfect covariation. We only used constructions that included asymmetrical features that always confirmed Mass theory but always disconfirmed Center theory. (2) We assessed children's prior theories at pretest and tried to confront them with evidence supporting Mass theory and contradicting their Center theory. (3) The children were free to play with the provided materials for an hour so that they had a sufficient amount of time to explore and play with the materials. Furthermore, we considered the children's developmental constraints and related science to their everyday activities by using different playful activities with building blocks as a learning setting (Copple and Bredenkamp, 2009).

The activities the children engaged in fulfilled the characteristics of play (Rubin et al., 1983; Pellegrini, 2013; Weisberg et al., 2013; Trawick-Smith, 2015; Daubert et al., 2018). The children played voluntarily, and the play was child-directed and contained elements of choice. We did not measure the children's motivation during the play. Therefore, we cannot make a statement about their intrinsic motivation. An indication of their motivation might be that the children could stop playing at any time, but approximately 95% of the children continued to play for the provided time in all groups, as the video recordings for the manipulation check and the experimenters' written records showed. Furthermore, highly inferential analyses of the recordings demonstrated that all of the children in all of the groups showed playful behavior (cf. Bundy et al., 2001). The play was not free of goals because we had a specific learning goal, namely, the acquisition of Mass theory, in mind. However, the play was still process-oriented, as we did not push this goal on the children.

Our playful intervention was based on the continual view postulated by, e.g., Zosh et al. (2018). The free play was free of an adult's guidance, the guided play with material scaffolds was structured and offered children suggestions for playing with blocks, and the guided play with material and verbal scaffolds offered additional verbal guidance. Specifically, when implementing the verbal scaffolds, we *asked for the children's prior knowledge* to allow them to express their presumptions to facilitate the adjustment of their theory (Mayer, 1997; Weinert and Helmke, 1998; Gurlitt and Renkl, 2010). By *asking for reasoning*, the children could be made aware of their theory, and they were supported in structuring their theory (Hsin and Wu, 2011). The *provision of explanations* helped the children organize new knowledge, e.g., knowledge about the mass, and integrate this new knowledge into their theories (Richey and Nokes-Malach, 2013). By *encouraging comparisons*, we tried to support the children in comparing stable and unstable constructions and to generalize the underlying principle, i.e., the mass (Hsin and Wu, 2011; Richey and Nokes-Malach, 2013). Last, *modeling* might have offered the children the possibility for imitation (Mayer, 2004; Hmelo-Silver et al., 2007).

Our study showed that crystallized intelligence had a positive relationship with children's consistent application of Mass theory directly after the intervention. The interaction of crystallized intelligence with the intervention group showed that children with high crystallized intelligence profited more in the Verbal group than did children with high crystallized intelligence in the Material group. Fluid intelligence did not relate to the consistent explanation with Mass theory directly after the intervention. This outcome indicates that children acquired Mass theory regardless of their ability to mentally represent and abstract the spatial features of constructions' stabilities. After including the follow-up, both fluid intelligence and crystallized intelligence related to the consistent explanation with Mass theory over an extended period of time. Crystallized intelligence interacted with the intervention group and related to the consistent application of

Mass theory in the Verbal group and the Free play group but not in the Material group. Children with low crystallized intelligence were more likely to adjust their theories in the Material group, while children with high crystallized intelligence were more likely to adjust their theories in the Verbal group.

Language capacity is understood as an indicator of crystallized intelligence (Cattell, 1987; Flynn and Blair, 2013). Thus, our results suggest that the children with a low language capacity profited most from the Material group. Seemingly, the material scaffolds were sufficiently self-explaining so that the children with a low language capacity could observe evidence for Mass theory and adjust their theories. Moreover, the children with a low language capacity in the Verbal group did not profit from the intervention because they may have suffered from a high cognitive load (Kirschner, 2002). They not only needed to process the new information about the learning content provided through the verbal scaffolds but also the language itself. In contrast, the children with a high language capacity profited from the verbal scaffolds that were provided in the Verbal group. Our findings imply that when providing verbal instructions and support, it may be important to consider children's language capacity.

Children with a higher fluid intelligence, i.e., a capacity to represent constructions mentally and abstract important spatial features (Cattell, 1987; Weiß and Osterland, 2013; Weber and Leuchter, 2020), were more likely to acquire Mass theory over the course of 10 weeks. This outcome is in line with studies from developmental psychology showing that children's Mass theory develops between ages five and seven (Bonawitz et al., 2012). In their everyday lives, children have many possibilities to explore stabilities and develop an understanding of the underlying principles. A possible explanation might be that children with a higher fluid intelligence learn these principles faster than children with a lower fluid intelligence (cf. Weber and Leuchter, 2020).

In addition to intelligence, other individual competencies, such as children's prior theories about stabilities, could relate to children's acquisition of Mass theory.

Relationship of Children's Prior Theories and Their Consistent Use of Mass Theory

The third research question was concerned with the role of children's prior theories on their consistent use of Mass theory after the interventions.

The children with inconsistent prior theories who received the highest amount of support (Verbal group) acquired a Mass theory, while those children who received less support (Material group and Free play group) did not acquire a Mass theory. This result indicates that prior theories play a role in theory adjustment, which is in line with findings concerned with Bayesian inference in the context of theory-theory by, e.g., Bonawitz et al. (2012) and Sobel et al. (2004). The children who answered inconsistently at T1 did not have a consistent prior theory and therefore had the highest chance of acquiring a Mass theory. Their theoretical assumptions were inconsistent compared to the theories of children who had explained their reasoning consistently with Center theory or Other (cf. Koerber et al., 2005; Griffiths et al., 2011; Gopnik and Wellman, 2012; Gopnik, 2013). Since the children in the Verbal group were most likely to acquire a Mass theory, children seem to profit from high amounts of guidance and support; moreover, in our study, only observation seems to be insufficient for the understanding of counterevidence. In sum, children with inconsistent prior theoretical assumptions profit from supported play but do not adjust their theories by playing with blocks freely.

Transfer Test

The fourth research question was concerned with children's performance on a transfer test at follow-up. Independent of intervention group, we compared children who had used Mass theory consistently after the playful intervention to children who had failed to do so on a transfer test with asymmetrical block constructions. The children who had answered consistently outperformed children who had answered inconsistently. This result indicates that children who explain their reasoning with Mass theory are also more likely to rate asymmetrical constructions' stabilities correctly, which suggests that the children acquired an understanding of Mass theory.

However, comparing the three intervention groups, children performed equally well on the transfer test, even though we found group differences on the reasoning test. The transfer test, unlike the reasoning test, was a paper-pencil test and according to Karmiloff-Smith (1992) and Pine and Messer (2003) rather tested knowledge that children might not have been able to verbalize. The transfer test indicates that children in all groups had knowledge about stabilities at T3, but only the children in the Verbal group were able to verbalize their reasoning.

The children who had a low language capacity succeeded in the transfer test but not in the reasoning task. They did not have to explain their reasoning in the transfer test; they were only required to decide about the constructions' stabilities. Although we tried to consider a low language capacity in the reasoning task by counting specific gestures as indicators for Mass theory, e.g., pointing to the Mass, or Center theory, e.g., pointing at the middle, the transfer test was seemingly easier for the children to handle. This is especially meaningful for children with a different native language because these children might face challenges in explaining their reasoning adequately but might be able to show their knowledge about stability with a non-verbal test. Therefore, to offer children the opportunity to show their knowledge about science phenomena such as stability, methods that do not require the children to speak might be helpful.

Limitations

There are some limitations to this study concerning the implementation and measurement.

Implementation of Play

Regarding the implementation of the playful intervention, we compared material scaffolds, material + verbal scaffolds and free play regarding their effects on children's Mass theory. The effect of verbal scaffolds uncoupled from material scaffolds was not investigated. Future studies could implement a verbal scaffolds group by presenting children with the same unstructured

building blocks a free play group receives and adding verbal scaffolds. Moreover, the implementation of a baseline group not receiving any intervention would allow investigating whether free play has an effect on children's theory adjustment toward Mass theory compared to children's development.

We videotaped only some of the playful interventions for a manipulation check; as some children or their parents denied permission to videotape, some interventions were only audio-recorded. Moreover, for a few interventions, neither videos nor recordings exist due to technical failures with the equipment. Therefore, children's behavior during play cannot be analyzed, even though there might be interindividual differences in how children interacted with the experimenter and used the provided materials. For example, some children might have asked for help more often or might have built with the building blocks more actively, while others may have instead watched other children build. Furthermore, the materials provided in the guided play groups served as suggestions, and children in all groups were free to build other buildings. From the existing videos and recordings, we assume that the children in the guided play groups played the suggested activities and used the scaffolding materials. However, some children might have built at a higher pace and thus may have built more of the suggested structures than other children. Last, regarding children's behavior, the amount of time that the children spent playing on their own or with other children, their manipulation of and their conversations about the building blocks might have contributed to children's Mass theory after the intervention. These factors should be investigated in a future study.

In this study, we only used a limited set of verbal scaffolds and did not control for the verbal scaffolds' adaptability. However, the adaptability might have contributed to children's acquisition of Mass theory. Therefore, children's and experimenters' behavior during play should be investigated in the future.

Measures

The children received eight items showing different block constructions, and three asymmetrical items were used to assess children's theories about stability. The other symmetrical items were used to familiarize the children with the test and motivate them during testing because children find it easier to estimate symmetrical constructions' stabilities (Krist, 2010). These Center theory-compliant items might have led some children to adopt a Center theory instead of remaining in the Other category, even though the evidence for Center theory was imperfect. The results of this study show that although the children received these Center theory items, many still adopted Mass theory after the playful intervention. Future studies might benefit from the use of more items, which would also prolong the testing time, as more asymmetrical as well as symmetrical items would be needed. This addition could impact the children's attention capacity and their motivation to participate.

Children received feedback about the constructions' stabilities during testing because they built the construction and then removed the supporting black block to ascertain whether they had rated the stability correctly. Therefore, children had the opportunity to learn during testing, and the items were dependent on each other. As a result, we could not just sum up the items, and every item was considered a point in time. Thus, we used methods of risk-event analysis to analyze the data. Independent measurements would allow for different statistical approaches, e.g., statistical procedures that refer to the mean. Thus, in future studies, to achieve independent measures, children could not build constructions on their own but only rate and explain stabilities on the basis of photographs so that they do not receive feedback about stability.

Nevertheless, our study indicates that guided play can support young children's science learning. Differing degrees of scaffolding in guided play can be beneficial for helping children with different prerequisites adjust their theories when observing new evidence.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

AW and ML contributed to conception and design of the study. AW organized the database, performed the statistical analysis, and wrote the first draft of the manuscript. AW, TR, and ML wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This research was funded by the German Research Foundation (DFG), grant number 290497409.

ACKNOWLEDGMENTS

We wish to thank Ina Plöger for helping develop the scaffolding materials and our research assistants Laura Löwe, Hanna Fuhrmann, Kristin Hoffmeister, Rebecca Ott, and Aliena Eichhorn for helping with data acquisition, data input, transcribing, and coding of the interviews. Last but not least, we would like to thank all children, who participated in our research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01737/full#supplementary-material

REFERENCES

Akerson, V. L., Buck, G. A., Donnelly, L. A., Nargund-Joshi, V., and Weiland, I. S. (2011). The importance of teaching and learning nature of science in the early childhood years. *J. Sci. Educ. Technol.* 20, 537–549. doi: 10.1007/s10956-011-9312-5

Alboukadel, K., Kosinski, M., and Biecek, P. (2019). *Survminer: Drawing survival curves using ggplot2 (Version 0.4.6) [Computer software]*. Available at: https://CRAN.R-project.org/package=survminer.

Alfieri, L., Brooks, P. J., Aldrich, N. J., and Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *J. Educ. Psychol.* 103, 1–18. doi: 10.1037/a0021017

Baillargeon, R., and Hanko-Summers, S. (1990). Is the top object adequately supported by the bottom object? Young infants' understanding of support relations. *Cogn. Dev.* 5, 29–53. doi: 10.1016/0885-2014(90)90011-H

Bonawitz, E. B., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., and Schulz, L. (2011). The double-edged sword of pedagogy: instruction limits spontaneous exploration and discovery. *Cognition* 120, 322–330. doi: 10.1016/j.cognition.2010.10.001

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., and Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cogn. Psychol.* 64, 215–234. doi: 10.1016/j.cogpsych.2011.12.002

Borriello, G. A., and Liben, L. S. (2018). Encouraging maternal guidance of preschoolers' spatial thinking during block play. *Child Dev.* 89, 1209–1222. doi: 10.1111/cdev.12779

Bundy, A. C., Nelson, L., Metzger, M., and Bingaman, K. (2001). Validity and reliability of a test of playfulness. *Occup. Ther. J. Res.* 21, 276–292. doi: 10.1177/153944920102100405

Butts, D. P., Hofman, H. M., and Anderson, M. (1994). Is direct experience enough? A study of young children's views of sounds. *J. Elemen. Sci. Educ.* 6, 1–16. doi: 10.1007/BF03170646

Cattell, R. B. (1987). *Intelligence: It's Structure, Growth and Action*. Amsterdam: Elsevier.

Chen, Z., and Klahr, D. D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Dev.* 70, 1098–1120. doi: 10.1111/1467-8624.00081

Chin, C. (2007). Teacher questioning in science classrooms: approaches that stimulate productive thinking. *J. Res. Sci. Teach.* 44, 815–843. doi: 10.1002/tea.20171

Cook, C., Goodman, N. D., and Schulz, L. E. (2011). Where science starts: spontaneous experiments in preschoolers' exploratory play. *Cognition* 120, 341–349. doi: 10.1016/j.cognition.2011.03.003

Copple, C., and Bredenkamp, J. (eds) (2009). *Developmentally Appropriate Practice in Early Childhood Programs Serving Children From Birth Through Age 8*, 3rd Edn. Washington, D.C: National Association for the Education of Young Children.

Cremin, T., Glauert, E., Craft, A., Compton, A., and Stylianidou, F. (2015). Creative little scientists: exploring pedagogical synergies between inquiry-based and creative approaches in early years science. *Educ. 3-13* 43, 404–419. doi: 10.1080/03004279.2015.1020655

Daubert, E., Ramani, G. B., and Rubin, K. H. (2018). "Play-based learning and social development," in *Encyclopedia on Early Childhood Development*, eds M. Boivin, R. D. Peters, and R. Tremblay (Ottawa: Centre of Excellence for Early Childhood Development).

DeLoache, J. S. (2014). "Early development of the understanding of symbolic artifacts," in *Wiley-Blackwell Handbooks of Developmental Psychology. The Wiley-Blackwell Handbook of Childhood Cognitive Development*, 2nd Edn, ed. U. Goswami (Hoboken, NJ: Wiley-Blackwell), 312–336.

Dunn, L. M., Dunn, D. M., Lenhard, A., Lenhard, W., Segerer, R., and Suggate, S. (2015). *PPVT 4. Peabody Picture Vocabulary Test*. Bengaluru: Pearson Assessment.

Eshach, H., and Fried, M. N. (2005). Should science be taught in early childhood? *J. Sci. Educ. Technol.* 14, 315–336. doi: 10.1007/s10956-005-7198-9

Ferrara, K., Hirsh-Pasek, K., Newcombe, N. S., Golinkoff, R. M., and Lam, W. S. (2011). Block talk: spatial language during block play. *Mind BrainEduc.* 5, 143–151. doi: 10.1111/j.1751-228X.2011.01122.x

Fisher, K. R., Hirsh-Pasek, K., Golinkoff, R. M., Singer, D. G., and Berk, L. E. (2011). "Playing around in school: Implications for learning and educational policy," in *.), The Oxford Handbook of the Development of Play*, ed. A. D. Pellegrini (Oxford: Oxford University Press), 341–362.

Fisher, K. R., Hirsh-Pasek, K., Newcombe, N., and Golinkoff, R. M. (2013). Taking shape: supporting preschoolers' acquisition of geometric knowledge through guided play. *Child Dev.* 84, 1872–1878. doi: 10.1111/cdev.12091

Flynn, J. R., and Blair, C. (2013). "The history of intelligence: new spectacles for developmental psychology," in *The Oxford Handbook of Developmental Psychology: Vol. 1: Body and Mind*, ed. P. D. Zelazo (Oxford: Oxford University Press.), 765–790.

Gelman, R., and Brenneman, K. (2004). Science learning pathways for young children. *Early Child. Res. Q.* 19, 150–158.

Gopnik, A. (2013). "Causality," in *The Oxford Handbook of Developmental Psychology: Body and Mind*, Vol. 1, ed. P. D. Zelazo (Oxford: Oxford University Press), 628–650.

Gopnik, A., and Meltzoff, A. (1997). *Words, Thoughts, and Theories*. Cambridge, MA: MIT Press.

Gopnik, A., Sobel, D. M., Schulz, L. E., and Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Dev. Psychol.* 37, 620–629. doi: 10.1037//0012-1649.37.5.620

Gopnik, A., and Wellman, H. M. (2012). Reconstructing constructivism: causal models, bayesian learning mechanisms, and the theory theory. *Psychol,. Bull.* 138, 1085–1108. doi: 10.1037/a0028044

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., and Gopnik, A. (2011). Bayes and blickets: effects of knowledge on causal induction in children and adults. *Cog. Sci.* 35, 1407–1455. doi: 10.1111/j.1551-6709.2011.01203.x

Gurlitt, J., and Renkl, A. (2010). Prior knowledge activation: How different concept mapping tasks lead to substantial differences in cognitive processes, learning outcomes, and perceived self-efficacy. *Instr. Sci.* 38, 417–433. doi: 10.1007/s11251-008-9090-5

Haden, C. A. (2010). Talking about science in museums. *Child Dev. Perspect.* 4, 62–67. doi: 10.1111/j.1750-8606.2009.00119.x

Hadzigeorgiou, Y. (2002). A study of the development of the concept of mechanical stability in preschool children. *Res. Sci. Educ.* 32, 373–391. doi: 10.1023/A:1020801426075

Hawkins, J., Pea, R. D., Glick, J., and Scribner, S. (1984). "Merds that laugh don't like mushrooms": evidence for deductive reasoning by preschoolers. *Dev. Psychol.* 20, 584–594. doi: 10.1037/0012-1649.20.4.584

Hirsh-Pasek, K., Golinkoff, R. M., Berk, L. E., and Singer, D. G. (2009). *A Mandate for Playful Learning in Preschool. Presenting the Evidence*. Oxford: Oxford University Press.

Hmelo-Silver, C. E., Duncan, R. G., and Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: a response to Kirschner. Sweller, and Clark (2006). *Educ. Psychol.* 42, 99–107.

Hogan, K., and Pressley, M. (1997). "Becoming a scaffolder of students' learning," in *Scaffolding Student Learning*, eds K. Hogan and M. Pressley (Cambridge, MA: Brookline), 185–191.

Hsin, C. T., and Wu, H. K. (2011). Using scaffolding strategies to promote young children's scientific understandings of floating and sinking. *J. Sci. Educ. Technol.* 20, 656–666. doi: 10.1007/s10956-011-9310-7

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Kirschner, P. A. (2002). Cognitive load theory: implications of cognitive load theory on the design of learning. *Learn. Instr.* 12, 1–10.

Klahr, D., and Nigam, M. (2004). The equivalence of learning paths in early science instruction: effect of direct instruction and discovery learning. *Psychol. Sci.* 15, 661–667. doi: 10.1111/j.0956-7976.2004.00737.x

Klahr, D., Zimmerman, C., and Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science* 333, 971–975. doi: 10.1126/science.1204528

Kleickmann, T., Tröbst, S., Jonen, A., Vehmeyer, J., and Möller, K. (2016). The effects of expert scaffolding in elementary science professional development on teachers' beliefs and motivations, instructional practices, and student achievement. *J. Educ. Psychol.* 108, 21–42. doi: 10.1037/edu0000041

Koerber, S., Sodian, B., Thoermer, C., and Nett, U. (2005). Scientific reasoning in young children: preschoolers' ability to evaluate covariation evidence. *Swiss J. Psychol.* 64, 141–152. doi: 10.1024/1421-0185.64.3.141

Krist, H. (2010). Development of intuitions about support beyond infancy. *Dev. Psychol.* 46, 266–278. doi: 10.1037/a0018040

Krist, H., Atlas, C., Fischer, H., and Wiese, C. (2018). Development of basic intuitions about physical support during early childhood: evidence from a novel eye-tracking paradigm. *Q. J. Exp. Psychol.* 2:174702181773719. doi: 10.1177/1747021817737196

Krist, H., Horz, H., and Schönfeld, T. (2005). Children's block balancing revisited: no evidence for representational redescription. *Swiss J. Psychol.* 64, 183–193. doi: 10.1024/1421-0185.64.3.183

Krist, H., and Krüger, M. (2012). Towards a new method for bridging the gap between "smart" infants and "dumb" preschoolers. *Eur. J. Dev. Psychol.* 9, 631–637. doi: 10.1080/17405629.2012.657895

Lazonder, A. W., and Kamp, E. (2012). Bit by bit or all at once? Splitting up the inquiry task to promote children's scientific reasoning. *Learn. Instr.* 22, 458–464. doi: 10.1016/j.learninstruc.2012.05.005

Leuchter, M., and Naber, B. (2019). Studying children's knowledge base of one-sided levers as force amplifiers. *J. Res. Sci. Teach.* 56, 91–112. doi: 10.1002/tea.21470

Leuchter, M., Saalbach, H., and Hardy, I. (2014). Designing science learning in the first years of schooling. An intervention study with sequenced learning material on the topic of 'floating and sinking'. *Int. J. Sci. Educ.* 36, 1751–1771.

Mantzicopoulos, P., Patrick, H., and Samarapungavan, A. (2013). Science literacy in school and home contexts: kindergarteners' science achievement and motivation. *Cogn. Instr.* 31, 62–119. doi: 10.1080/07370008.2012.742087

Martin, N. D., Dornfeld Tissenbaum, C., Gnesdilow, D., and Puntambekar, S. (2019). Fading distributed scaffolds: the importance of complementarity between teacher and material scaffolds. *Instr. Sci.* 47, 69–98. doi: 10.1007/s11251-018-9474-0

Mayer, R. E. (1997). Multimedia learning: are we asking the right questions? *Educ. Psychol.* 32, 1–19. doi: 10.1207/s15326985ep3201_1

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *Am,. Psychol.* 59, 14–19. doi: 10.1037/0003-066X.59.1.14

Murphy, N., and Messer, D. (2000). Differential benefits from scaffolding and children working alone. *Educ. Psychol.* 20, 17–31. doi: 10.1080/014434100110353

Pellegrini, A. D. (2013). "Play," in *The Oxford Handbook of Developmental Psychology: Self and Other*, Vol. 2, ed. P. D. Zelazo (Oxford: Oxford Universiy Press), 276–299.

Pellegrini, A. D., and Gustafson, K. (2005). "Boys' and Girls' Uses of Objects for Exploration, Play, and Tools in Early Childhood,"in *The Nature of Play: Great Apes and Humans* eds A. D. Pellegrini and P. Smith (New York, NY: Guilford Press.), 113–138.

Piekny, J., and Maehler, C. (2013). Scientific reasoning in early and middle childhood: the development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *Br. J. Dev. Psychol.* 31(Pt 2), 153–179. doi: 10.1111/j.2044-835X.2012.02082.x

Pine, K. J., Lufkin, N., Kirk, E., and Messer, D. (2007). A microgenetic analysis of the relationship between speech and gesture in children: evidence for semantic and temporal asynchrony. *Lang. Cogn. Process.* 22, 234–246. doi: 10.1080/01690960600630881

Pine, K. J., and Messer, D. (2003). The development of representations as children learn about balancing. *Br. J. Dev. Psychol.* 21, 285–301.

Pine, K. J., Messer, D., and Godfrey, K. (1999). The teachability of children with naive theories: an exploration of the effects of two teaching methods. *Br. J./Educ. Psychol.* 69, 201–211.

Pyle, A., DeLuca, C., and Danniels, E. (2017). A scoping review of research on play-based pedagogies in kindergarten education. *Rev. Educ.* 5, 311–351. doi: 10.1002/rev3.3097

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Renkl, A. (2002). Worked-out examples: instructional explanations support learning by self-explanations. *Learn. Instr.* 12, 529–556. doi: 10.1016/S0959-4752(01)00030-5

Reuter, T., and Leuchter, M. (2020). Children's concepts of gears and their promotion through play. *J. Res. Sci. Teach.* doi: 10.1002/tea.21647 [Epub ahead of print].

Richey, J. E., and Nokes-Malach, T. J. (2013). How much is too much? Learning and motivation effects of adding instructional explanations to worked examples. *Learn. Instr.* 25, 104–124. doi: 10.1016/j.learninstruc.2012.11.006

Riley, W. F., and Sturges, L. D. (1993). *Engineering Mechanics: Statics*, 2nd Edn. Hoboken, NJ: Wiley.

Rubin, K. H., Fein, G. G., and Vandenberg, B. (1983). "Play," in *Handbook of Child Psychology*, ed. E. M. Hetherington (Hoboken, NJ: Wiley), 693–774.

Rubin, K. H., Watson, K. S., and Jambor, T. W. (1978). Free-Play behaviors in preschool and kindergarten children. *Child Dev.* 49, 534–536. doi: 10.2307/1128725

Ruffman, T., Perner, J., Olson, D. R., and Doherty, M. (1993). Reflecting on scientific thinking: children's understanding of the hypothesis-evidence relation. *Child Dev.* 64, 1617–1636. doi: 10.2307/1131459

Sarama, J., and Clements, D. H. (2009). "Concrete" computer manipulatives in mathematics education. *Child Dev. Perspect.* 3, 145–150. doi: 10.1111/j.1750-8606.2009.00095.x

Schulz, L. E., Bonawitz, E. B., and Griffiths, T. L. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Dev. Psychol.* 43, 1124–1139. doi: 10.1037/0012-1649.43.5.1124

Siegler, R. S. (1976). Three aspects of cognitive development. *Cogn. Psychol.* 8, 481–520.

Siegler, R. S., and Chen, Z. (1998). Developmental differences in rule learning: a microgenetic analyses. *Cogn. Psychol.* 36, 273–310.

Singer, J. D., and Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modelling Change and Event Occurence*. Oxford: Oxford University Press.

Sobel, D. M., Tenenbaum, J. B., and Gopnik, A. (2004). Children's causal inferences from indirect evidence: backwards blocking and Bayesian reasoning in preschoolers. *Cogn. Sci.* 28, 303–333. doi: 10.1016/j.cogsci.2003.11.001

Therneau, T. (2015). *A Package for Survival Analysis in S. (Version 2.38) [Computer software*]. Available at: https://CRAN.R-project.org/package=survival.

Thomson, D., Casey, B. M., Lombardi, C. M., and Nguyen, H. N. (2018). Quality of fathers' spatial concept support during block building predicts their daughters' early math skills – But not their sons'. *Early Child. Res. Q.* 50, 11–96. doi: 10.1016/j.ecresq.2018.07.008

Thorsen, C., Gustafsson, J. E., and Cliffordson, C. (2014). The influence of fluid and crystallized intelligence on the development of knowledge and skills. *Br. J. Educ. Psychol.* 84, 556–570. doi: 10.1111/bjep.12041

Trawick-Smith, J. (2015). "Teacher–child play interactions to achieve learning outcomes," in *Research in Early Childhood Science Education*, eds K. C. Trundle and M. Saçkes (Berlin: Springer), 259–277.

Trundle, K. C., and Saçkes, M. (2012). "Science and early education," in *Handbook of Early Childhood Education*, ed. R. Pianta (New York, NY: Guilford Publications), 240–258.

van de Pol, J., Volman, M., and Beishuizen, J. (2010). Scaffolding in teacher–student interaction: a decade of research. *Educ. Psychol. Rev.* 22, 271–296. doi: 10.1007/s10648-010-9127-6

van Schijndel, T. J. P., Visser, I., van Bers, B. M. C. W., and Raijmakers, M. E. J. (2015). Preschoolers perform more informative experiments after observing theory-violating evidence. *J. Exp. Child Psychol.* 131, 104–119. doi: 10.1016/j.jecp.2014.11.008

Verdine, B. N., Zimmermann, L., Foster, L., Marzouk, M. A., Golinkoff, R. M., Hirsh-Pasek, K., et al. (2019). Effects of geometric toy design on parent-child interactions and spatial language. *Early Child. Res. Q.* 46, 126–141. doi: 10.1016/j.ecresq.2018.03.015

Weber, A. M., and Leuchter, M. (2020). Measuring preschool children's knowledge of the principle of static equilibrium in the context of building blocks: validation of a test instrument. *Br. J. Educ. Psychol.* 90, 50–74. doi: 10.1111/bjep.12304

Weinert, F. E., and Helmke, A. (1998). The neglected role of individual differences in theoretical models of cognitive development. *Learn. Instr.* 8, 309–323. doi: 10.1016/S0959-4752(97)00024-8

Weisberg, D. S., Hirsh-Pasek, K., Golinkoff, R. M., Kittredge, A. K., and Klahr, D. (2016). Guided play. *Curr. Dir. Psychol. Sci.* 25, 177–182. doi: 10.1177/0963721416645512

Weisberg, D. S., Zosh, J. M., Hirsh-Pasek, K., and Golinkoff, R. M. (2013). Talking it up: play, language development, and the role of adult support. *Am. J. Play* 6, 39–54.

Weiß, R. H., and Osterland, J. (2013). *CFT 1-R*. Boston, MA: Hogrefe.

Zosh, J. M., Hirsh-Pasek, K., Hopkins, E. J., Jensen, H., Liu, C., Neale, D., et al. (2018). Accessing the inaccessible: redefining play as a spectrum. *Front. Psychol.* 9:1124. doi: 10.3389/fpsyg.2018.01124

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Weber, Reuter and Leuchter. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 July 2020
doi: 10.3389/fpsyg.2020.01790

Nascent Inquiry, Metacognitive, and Self-Regulation Capabilities Among Preschoolers During Scientific Exploration

*Ronit Fridman[1,2], Sigal Eden[1,2] and Ornit Spektor-Levy[2]**

[1] Faculty of Social Sciences, School of Education, Bar-Ilan University, Ramat Gan, Israel, [2] The School of Education, Bar Ilan University, Ramat Gan, Israel

There is common agreement that preschool-level science education affects children's curiosity, their positive approach toward science, and their desire to engage with the subject. Children's natural curiosity drives them to engage enthusiastically in all forms of exploration. Engaging in scientific exploration necessitates self-regulation capabilities and a wide repertoire of cognitive and metacognitive strategies. The purpose of this study was to examine to what extent preschoolers (aged 5–6 years) implement nascent inquiry skills, metacognitive awareness, and self-regulation capabilities during play-based scientific exploration tasks. An additional purpose was to investigate the relationships between these capabilities, a relationship not yet investigated in the context of play-based, scientific exploration among young children. The study consisted of 215 preschoolers, from 10 preschools. For this study, we developed two scientific exploration tasks – structured and open-ended. Our motivation was to examine whether preschoolers' capabilities will differ in the context of structured task which is aligned with the view that young children need guidance and explicit instructions compared to the context of open-ended, play-based task–allowing the children to apply and test their intuitive theories and skills. During performance participants were videotaped. Their verbal and non-verbal responses were analyzed by means of a coding scheme. The results of a micro-analysis of about 100 h of video showed that given the opportunity, even without setting explicit goals and instructions, children exhibit inquiry capabilities: they ask questions, plan, hypothesize, use tools, draw conclusions. Asking questions and planning were better manifested during the structured task. Children also manifested higher levels of attention, persistence, and autonomy during the structured task. However, significant higher scores of self-regulation indications were revealed in the context of the open-ended, play-based, exploration task. Moreover, results indicate significant correlations between the five measures of preschoolers' inquiry capabilities and measures of metacognitive strategic awareness and self-regulation. The results of the present study suggest the importance of combining various learning environments and experiences in early science education that encourage children to engage in structured exploration alongside play-based, open-ended, exploration.

Keywords: early childhood, science education, inquiry, metacognition, self-regulation, exploration

OPEN ACCESS

Edited by:
Christopher Osterhaus, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Deborah Sonia Pino-Pasternak, University of Canberra, Australia
Kaichi Yanaoka, The University of Tokyo, Japan
Jos Beishuizen, Vrije Universiteit Amsterdam, Netherlands

***Correspondence:**
Ornit Spektor-Levy
ornit.spektor-levy@biu.ac.il

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *28 February 2020*
Accepted: *29 June 2020*
Published: *21 July 2020*

Citation:
Fridman R, Eden S and Spektor-Levy O (2020) Nascent Inquiry, Metacognitive, and Self-Regulation Capabilities Among Preschoolers During Scientific Exploration. Front. Psychol. 11:1790. doi: 10.3389/fpsyg.2020.01790

INTRODUCTION

There is a consensus among researchers that preschool-level science education affects children's curiosity, their positive approach toward science, and their desire to engage with the subject (Eshach, 2006; Patrick et al., 2009; Worth, 2019). These factors predict the likelihood of engaging in science and of scientific achievements in both the short and long terms (Osborne et al., 2003; Tao et al., 2012). The American National Science Teachers Association (National Science Teachers Association [NSTA], 2014) recommended the adoption of an official position on teaching science in early childhood, claiming that the scientific inquiry process is based on principles of active and independent activity on the part of the child. The aspiration is to bring the child to a level of open, independent inquiry, wherein s/he can raise research queries on her/his own, plan the research, and carry it out. In addition, the National Science Teachers Association [NSTA] (2014) presented principles aimed at shedding light on those activities, and ways to make engagement in science accessible to children to help them develop skills and knowledge over time. As engagement in science is a knowledge-building process, it is recommended to expose children to many varied opportunities to engage in scientific inquiry processes on a regular basis. Children need time to play, to observe new phenomena, to think about what they have seen and discovered, and to draw conclusions (National Science Teachers Association [NSTA], 2014). These complex capabilities require learners to criticize and regulate their own thinking, learning, and outcomes. Thus, the scope of this study involves identifying these nascent capabilities in order to nurture and develop them through providing inquiry-based learning opportunities.

LITERATURE REVIEW

Early Childhood Science Education

Over the years, scholarly literature has increasingly acknowledged the advantages of beginning science education in early childhood (Eshach, 2003; National Science Teachers Association [NSTA], 2014; Early Childhood Stem Working Group, 2017; McClure et al., 2017).

One important objective of science education is to assist children in discovering the world and to answer their questions as they employ their cognitive and physical skills (Jones et al., 2008). It is essential to present science to children in developmentally appropriate ways, thus enabling them to explore the world through sensory investigations. This also helps children to absorb basic knowledge and abilities that are necessary for lifelong science learning and appreciation of nature (Trundle and Saçkes, 2015). The guided process could begin by observing children during their explorations and offering them appropriate support to improve their thinking and inquiries.

Some researchers view the child as a natural scientist (Gopnik, 2012), a contention based on studies indicating that young children possess cognitive abilities enabling them to comprehend scientific concepts and implement inquiry skills (Buchsbaum et al., 2011). According to this view, children understand and explain the world by means of intuitive theories and are able to transform these theories in light of new, cumulative knowledge (Williams and Lombrozo, 2013). In contrast, other studies argue that children have difficulty spontaneously acquiring the components of scientific inquiry and implementing scientific thinking skills – such as designing experiments, recording data, and processing existing knowledge with new knowledge – without guidance and explicit instruction. Therefore, they must be taught all these capabilities in a structured way (Zimmerman, 2007).

Inquiry in Preschool

Implementation of early childhood science curricula and exposing children to enjoyable, exciting science activities can help them develop scientific knowledge, inquiry capabilities, scientific discourse, and positive attitudes toward science (French, 2004; Samarapungavan et al., 2008; Patrick et al., 2009). Policy papers and scholarly literature have recommended teaching science in the manner conducted by scientists; that is, via inquiry (e.g., National Research Council (NRC), 2012; NGSS Lead States, 2013). Inquiry-based learning reflects the constructivist approach, according to which educators must strive to create a learning environment wherein the learners are required to examine thought processes: to gather, record, and analyze data; to analyze and test hypotheses; to test prior knowledge, and to formulate new significance on their own (Dewey, 1938; Kuhn et al., 2000; Greeno, 2002; Loyens and Rikers, 2011; Andiema, 2016). Indeed, young children can hypothesize and modify hypotheses as needed (Legare, 2012); provide explanations that draw upon high-level, abstract thinking; and process existing knowledge with new knowledge (Schulz, 2012; Williams and Lombrozo, 2013). They can carry out observations, establish hypotheses based on evidence, understand experiments, identify reliable information sources (Klahr et al., 2011), use their own hypotheses to predict results, and evaluate evidence (Piekny and Maehler, 2013; Walker et al., 2014).

It is recommended that children be engaged in science topics, and that science constantly be incorporated into other subjects learned in preschool (National Research Council (NRC), 2012; NGSS Lead States, 2013). The first step in teaching the sciences is to enable children to engage with various objects and materials and to provide a range of possibilities, materials, and opportunities (Sheridan and Samuelsson, 2001). It is fundamental to direct any activity with suitable inquiries and practices, rather than tell the children what to do (Tunnicliffe, 2015). The engagement with scientific thinking skills in such varied ways provides children with opportunities to develop other skills, such as mathematical language and social skills (Bustamante et al., 2018). Moreover, science learning and the scientific inquiry process provide stimulating contexts for the development of metacognitive and self-regulation capabilities (Michalsky et al., 2007; Zimmerman, 2008; Jirout and Zimmerman, 2015).

Metacognition and Self-Regulation

In recent years, the term *metacognition* has been widely used in the field of education and the study of metacognition has evolved into a flourishing field of research in cognitive psychology. Flavell (1976), who coined the term metacognition, defined it as the individual's knowledge, regulation, and control of the processes and outcomes of her/his own cognitive system. Over the years, this definition has expanded to include the emergence of associated concepts such as reflection, self-regulation, metacognitive awareness, etc. Metacognition is higher-order thinking, and includes critical thinking about one's own thinking, planning, controlling, monitoring, assessing, knowing what information is required for a task, and how to use the appropriate tools needed to perform the task (Flavell, 1979).

The increased number of studies of the concept of metacognition has led to confusion and discussions about the distinctions between the various components of metacognition and their relationships (Veenman et al., 2006). The scope of this review does not allow for a survey of all the approaches and taxonomies of metacognition.

We have adopted the view that distinguishes between two major components of metacognition: knowledge about cognition and regulation of cognition (Flavell, 1976; Brown, 1987; Schraw and Dennison, 1994). Knowledge about cognition is realized through three reflective processes: declarative knowledge (what), procedural knowledge (how), and conditional knowledge (when and why). Regulation of cognition is realized through the following processes: planning, monitoring, and management of strategies – control, debugging, and evaluation (Schraw and Dennison, 1994; Schraw and Moshman, 1995).

In this study, regulation of cognition is called *self-regulation*. Self-regulation refers to how we use external and internal clues to determine when to start, continue, or stop a particular action to achieve the desired goal. Accordingly, self-regulation includes the ability to choose behaviors while controlling the intensity of the response, the ability to plan the response, and to respond effectively during internal and interpersonal discourse (Zimmerman, 2000).

Self-regulation is one indicator of the relationship between emotion and cognition (Lee, 2005). A wide range of motivation-related cognitive interactions and metacognition control the process of acquiring self-regulation skills (Zimmerman, 2000; Schraw et al., 2006). Motivation relates to perceptions, attitudes, and desires that influence the use and development of cognitive and metacognitive skills (Schraw et al., 2006).

At every age, self-regulation for learning is the ability to identify or set goals; identify a mismatch between goals and the state of one's current expertise; monitor learning behaviors continuously and accurately; and initiate regulation processes for the performance of a task (Best and Miller, 2010).

Metacognition and Self-Regulation Among Young Children

According to Vygotsky (1978, 1986), learning begins in a social context, where an adult supports the young learner, providing her/him with a safety net. It is a process of internalization that begins with regulation through others and evolves into self-regulation.

Evidence of self-regulation of learning and cognition in early childhood has accumulated in recent decades (e.g., Istomina, 1975; Schweinhart and Weikart, 1997; Bronson, 2000; Whitebread et al., 2005, 2007, 2009; Blair and Razza, 2007; Rodríguez and Palacios, 2007; Hattie, 2009). In addition, much research evidence about metacognitive knowledge and self-regulation has accumulated regarding children of 3–5 years of age (Mevarech, 1995; Shamir et al., 2009; Whitebread et al., 2009).

Studies such as the longitudinal national cohort study, Pre-COOL (Mulder et al., 2014), and the Effective Provision of Pre-School Education (EPPE) project (Sylva et al., 2004) have indicated the importance of self-regulated learning in early childhood. Fostering self-regulated learning as early as possible is crucially important, since children develop their learning abilities during their first early years (De Corte et al., 2000; Hendy and Whitebread, 2000). Bryce and Whitebread (2012), as well as Bryce et al. (2015), contend that monitoring and control are some of the abilities that have already been developed by preschool age. Therefore, it seems prudent to foster young children's metacognitive and self-regulation capabilities at an early stage (Winne and Hadwin, 2008). Fostering these capacities proves beneficial to scholastic performance (Butler et al., 2004; Blair and Razza, 2007; Hidi and Ainley, 2008; Rimm-Kaufman et al., 2009; Moffitt et al., 2011; Butler and Schnellert, 2012; Diamond et al., 2013; Dunn et al., 2014; Kitsantas et al., 2017; Dörr and Perels, 2019).

Metacognitive thinking develops intuitively in children, along with the evolution of intelligence through their interaction with the environment, with parents, with teachers, with friends, and with others (Beishuizen and Veenman, 2004). According to Bronson (2000), the optimal environment for encouraging the development of emotional and behavioral self-regulation has a number of characteristics: It is regular, safe, sufficiently stimulating, responsive, and sensitive to the child's needs and perspectives, and affords security and encouragement. It sets clear boundaries and standards, offer examples and role modeling, and provides opportunities for activities that are directed toward the development of autonomous self-regulation (Klein and Yablon, 2008).

In this study we will focus on nascent metacognitive thinking in the context of play-based scientific experiments among young children. Since it is challenging to differentiate between the three components comprising knowledge of cognition: (declarative, procedural, and conditional; Schraw and Dennison, 1994) in very young children, we will relate to the general manifestation of metacognitive strategic awareness (e.g., looking for evidence like planning, looking at the available materials, and pausing to think). We will also look for evidence of nascent self-regulation (e.g., task awareness, planning, monitoring and debugging, controlling, and evaluating).

Inquiry and Self-Regulation Capabilities

Over the past three decades, research on cognitive development has yielded data on the development of children as independent learners. Inquiry and exploration processes provide opportunities

to develop independent learning capabilities. During the course of inquiry processes, a process of formulating knowledge and, when needed, transforming knowledge takes place, and more sophisticated inquiry skills are developed. Concurrently, as metacognitive and meta-strategic knowledge develop, children and adolescents gain a better understanding of the nature of inquiry and the skills used in the inquiry process. Thus, the entire process is iterative and cyclical, involving some or all of the components of scientific inquiry – such as designing experiments, evaluating evidence, and drawing inferences – serving to form and/or revise theories about the phenomena under investigation (Jirout and Zimmerman, 2015). Kuhn (1989, 2002) contends that the defining feature of scientific thinking is the set of cognitive and metacognitive skills involved in differentiating and coordinating theory and evidence. In particular, metacognitive awareness is what differentiates more sophisticated scientific thinking from less sophisticated. Instruction that develops higher-order thinking contributes to building students' knowledge and aids their transition from memorizing and rote learning to learning that emphasizes the building of knowledge in meaningful ways. It encourages their self-regulation and motivation to succeed at science by means of modeling effective inquiry strategies (Chinn and Malhotra, 2002). Therefore, understanding the ways nascent inquiry and self-regulation capabilities emerge, develop and are mutually related in early childhood is a prerequisite.

A review of the literature on early childhood science education reveals that despite the existing volume of research accumulated over the years, what we know about very young children's science learning is limited, as most of the research has focused on the later elementary school years and beyond (Trundle and Saçkes, 2015). While this may be partially due to a lack of priority in early childhood research funding, it is also important to acknowledge the challenges of researching scientific knowledge and understanding among very young children, where language, both oral and written, may not reflect children's skills, reasoning, and understanding (National Science Teachers Association [NSTA], 2014). Therefore, the main aim of this study is to analyze verbal and non-verbal responses of preschool children (5–6 years of age), to identify their nascent and intuitive inquiry thinking and behaviors, metacognitive strategic awareness, self-regulation capabilities and the relation between these capabilities, while they are engaged in structured and open-ended, play-based scientific experiences. Previous studies, as noted above, showed significant evidence that engaging in scientific inquiry processes necessitates self-regulation capabilities and a wide repertoire of cognitive and metacognitive strategic awareness. The extent to which these combined capabilities are developed and manifested in early childhood are rarely investigated among preschool children (5–6 years of age), particularly in the context of open-ended scientific exploration.

Play-Based Scientific Exploration

A great deal of research has focused on testing and describing children's responses while encountering a novel occurrence in a structured setting and inferring the children's assumptions and conceptions. Less attention has been paid to the ways that these processes occur in everyday environments and the ways these processes occur in free-play and play-like situations (Wager and Parks, 2014). Why play? Researchers and educators are often skeptical of play-based learning activities, seeing them as just play. However, as Bergen (2009) contends, playfulness appears to provide a predisposition toward certain types of creative acts, including those employed in scientific and mathematical fields. Play is valuable for children primarily because it is a medium for development and learning (Bergen, 2009). Play enables children to examine materials and try out techniques in artistic and creative endeavors. Therefore, scientific exploration tasks may provide play-based situations in which to examine preschoolers' intuitive inquiry and self-regulation processes.

In this present study, we designed a mini Manipulative Environment (ME; Bumbacher et al., 2018) that invites preschoolers to explore and manipulate materials using an inquiry-based approach. The representational aspects of the ME relate to the concept of affordances: the information a ME provides which builds the basis for decisions of how to interact with it.

While planning the ME, we were inspired by Fleer (2009) who described a "potion center" (p. 1076) to explore materials and their properties. The potion-play involved the provision of numerous plastic bottles, plastic tubing, buckets, colored water, and funnels. In addition, at appointed times the children were also given fragrant oils, vinegar, peanut oil, and more. Following Fleer (2009), we designed two mini MEs for play-based scientific exploration implementing both open-ended and structured play-like tasks.

The play-based scientific exploration tasks developed for this study were performed in preschool classrooms. The open-ended tasks did not impose clear task goals for the child to reach. This authentic, ecological methodology has rarely been applied so far, and thus has the potential to contribute to the theoretical and practical knowledge in the field of early science education.

RESEARCH QUESTIONS

The purpose of this study is to examine to what extent preschoolers implement nascent intuitive inquiry processes, metacognitive strategic awareness, and self-regulation capabilities during open-ended and structured play-based science experiences, and to what extent a relationship exists between these capabilities. Such relationships have not yet been studied among young children in the context of an authentic, ecological approach involving play-based science experiences in the preschool classroom.

We gathered data directly from the children during the performance of two scientific exploration tasks. Our three research questions and hypotheses were as follows:

1. To what extent do preschoolers manifest intuitive inquiry capabilities during open-ended and structured play-based scientific exploration tasks?

In this present study, we hypothesize that in the context of play, children will perform intuitive inquiry processes to

a significant extent. This hypothesis is based on structured experiments showing that young children can carry out observations, establish hypotheses based on evidence, interpret, and infer (Klahr et al., 2011; Piekny and Maehler, 2013; Walker et al., 2014). Moreover, Bergen (2009) argues that free play is valuable for children and provides an enjoyable medium for development and learning. Play enables children to examine materials and try techniques in explorative and creative endeavors. Therefore, our hypothesis will be tested in an open-ended, play-based scientific exploration task as well as in a structured task.

2. To what extent do preschoolers manifest metacognitive strategic awareness and self-regulation capabilities during open-ended and structured play-based scientific exploration tasks?

We hypothesize that in the context of play, preschoolers will manifest similar significant metacognitive strategic awareness and self-regulation capabilities as previously noted in the Literature Review, whereby there is accumulated evidence of preschoolers' capabilities to perform metacognitive awareness and self-regulation capabilities in the context of structured tests (Bronson, 2000). In a previous study, young children showed such capabilities during play-based construction tasks (Spektor-Levy et al., 2017). Therefore, our hypothesis will be tested in an open-ended, play-based scientific exploration task as well as in a structured task.

3. What is the nature of the relationship between the implementation of inquiry processes, metacognitive strategic awareness, and self-regulation capabilities among preschoolers during an open-ended and structured play-based scientific exploration task?

We hypothesize that positive, medium-strength correlations will be found between the implementation of inquiry processes, metacognitive awareness, and self-regulation capabilities among preschoolers during an open-ended and structured play-based scientific exploration task, as the literature is replete with such evidence among older students (Kuhn, 1989, 2002; Jirout and Zimmerman, 2015). Following a thorough search of the scholarly literature, we scarcely found any studies that examined the relations between metacognitive strategic awareness, self-regulation capabilities and inquiry capabilities among preschoolers during authentic play-based scientific exploration in the context of scientific exploration tasks.

METHODOLOGY

Study Sample

The study sample consisted of 215 children: 120 boys (55.8%) and 95 girls (44.2%). The mean age of the participants was 64.79 months (*SD* = 4.2). Children were randomly assigned from 10 urban, middle SES, mainstream preschool classrooms, with parents' consent. The study was reviewed and approved by the Ethics Committee #10495, The Office of the Chief Scientist of the Ministry of Education, Israel.

Baseline measures were collected using two tools: The *Raven's Progressive Matrices* (Raven, 1956), designed to measure general cognitive ability and meaning-making. Raven scores ranged from 85 to 129 (M = 110.07; SD = 10.75). The *Peabody Picture Vocabulary Test* (PPVT-4; Dunn and Dunn, 2007), which measures an individual's receptive vocabulary, providing an estimate of verbal ability and scholastic aptitude. PPVT scores ranged from 85 to 128 (M = 109.43; SD = 8.38). All participants scored within the typical range. None of the participants were diagnosed as having any developmental or language delays, nor any motor disabilities.

In Israel, the preschool (3–6 years of age) curriculum requires the implementation of an obligatory Science and Technology curriculum that emphasizes inquiry activities. However, all teachers of the 10 preschool classrooms in this study reported that they employ neither open inquiry processes nor the acquisition of inquiry practices in their classrooms.

Data collection was carried out over a 2-year period. During the first year, data were collected from 68 children who engaged in the free, open-ended scientific exploration task. During the second year, data were collected from 147 children who engaged in the open-ended scientific task and another task, a structured scientific exploration task. For the entire sample of 215 preschoolers, only descriptive statistics and correlations were used.

The optimal overall sample size was determined a *priori* using G Power software. The desired sample size is at least 197 subjects for a two-tailed hypothesis, a small effect size of 0.25, an alpha 0.05 error, and a very high power of 0.95.

The optimal sample size of the 147 subjects, for whom Wilcoxon tests were performed to analyze paired comparisons, was also determined *a priori* using G Power software. The desired sample size is at least 146 subjects for a two-tailed hypothesis, a small effect size of 0.25, an alpha 0.05 error, and a high power of 0.85.

Research Procedure

Data collection employed a mixed-method approach, combining qualitative and quantitative research tools. In a mixed method study, the researcher or group of researchers combine elements from the qualitative and quantitative research methods in order to expand, deepen, and reinforce the intellectual and practical insights based on quantitative and qualitative evidence (Johnson et al., 2007).

Task performance was carried out with each child individually, during school hours, in a quiet space within the preschool facility. Each task included a pre- and post-exploration interview. The first phase of the pre-exploration interview was intended to facilitate the acquaintance between the researcher and the child, and to gather background information (e.g., "What's your name?," "How old are you?"). The next phase of the pre-exploration interview included the following questions: "Do you like to investigate things? If you do like to investigate things, how do you go about it?," "What's on the tray?," "What would you like to do with the materials and equipment on the tray?," "What do you think will happen?," "Do you have any questions? Would you like to know more about anything that's here?" After

the exploration task, a post-interview was conducted: "What happened during your investigation?," "Why do you think it happened? Can you explain?," "Are you satisfied/happy with what happened? Why?," "If you had more time, what else would you have done?," "Do you like to investigate things like what we did here?". Next, the two scientific exploration tasks were performed in a counterbalanced manner.

The scientific exploration tasks afforded opportunities for planning, hypothesizing, concluding, asking questions, using scientific tools, self-regulation, and constant awareness of task requirements, monitoring, and problem-solving capabilities involving trial and error.

Research Tools

Exploration Tasks

As young children lack the verbal proficiency necessary for prospective and retrospective self-report measures (e.g., questionnaires, interviews, and thinking aloud), it is increasingly recognized that research with very young children should be based on behavioral, exploratory methodologies (Winne and Perry, 2000; Whitebread et al., 2005). Studies relying less upon children's verbal abilities have tended to show children to be more knowledgeable and skilled than originally believed (Whitebread et al., 2009; Spektor-Levy et al., 2017). Therefore, for this study, we developed two tasks to identify indications of nascent inquiry, metacognitive strategic awareness, and self-regulation capabilities during play-based scientific exploration.

Both tasks lasted about 15 min and comprised three parts: a pre-exploration interview, a scientific exploration task (open-ended or structured), and a post-exploration interview. In the pre-exploration interview, the participants were asked about their attitudes toward science, what they knew about science, and when and how they had engaged in science. In addition, both interviews (pre- and post-exploration) included questions aimed at examining and identifying the nascent scientific thinking of the children, and their ability to implement components of inquiry such as asking questions, planning, hypothesizing, and drawing conclusions (see **Supplementary Data Sheet 1**). The interviews and the exploration took place during preschool hours, and each child was interviewed individually spoons (**Figure 1**).

Throughout the exploration tasks, video-recordings were made of the children's behavior and responses. Our intention was to closely observe and record the verbal and non-verbal-behavioral responses of the preschoolers while they were intuitively engaged in the exploration tasks. Micro-analysis of the video recordings (described below in the coding scheme) enabled the researchers to quantify the inquiry, metacognitive strategic awareness, and self-regulation capabilities that were identified.

Open-ended scientific exploration task (N = 215)

On the table there was a tray with jars containing various liquids (still water, seltzer water, oil, juice, and milk) and the following items: a syringe, two droppers, a funnel, measuring cups, and measuring spoons. The children were asked to experiment freely with the liquids and objects for 5 min. They were told, "You have 5 min to do whatever you want with these items. When the timer goes off, your time is up, and you have to finish."

Structured exploration task (N = 147)

This task followed the Predict-Observe-Explain (POE) approach, which requires that before each experiment learners make predictions about the outcome (Bumbacher et al., 2018). Then they design the experiment and re-examine the predictions in light of their new observations (Rickey and Stacy, 2000). Despite its caveats, POE is a simple intervention that prompts learners to think more carefully about the experiment's design, expected outcomes, and what can be learned, which manifests intentional and unconfounded experimentation capabilities (Kearney et al., 2001).

In this task, the table held a tray with jars containing liquids (two jars containing still water, one jar containing juice, one containing oil, two empty jars), a syringe, two droppers, a funnel, measuring cups, and measuring spoons (**Figure 1**). The children were asked to follow these instructions: "Mix half of the water with half of the oil, and half of the water with half of the juice; and then tell us when you're done. After telling us you're done, leave the items in place."

Coding Scheme

The coding scheme was based on the Metacognitive Skills in Constructional Play Engagement (MetaSCoPE) coding scheme. It was designed by Bryce and Whitebread (2012) and was further developed for accuracy by Spektor-Levy et al. (2017). For the present study, the coding scheme was adapted to identify emerging inquiry capabilities. The video analysis coded both verbal and non-verbal responses. Non-verbal responses could be manifested by private gestures, that is, signs that children intentionally direct toward themselves or objects (Rodríguez and Palacios, 2007; Basilio and Rodríguez, 2011). Private gestures may reflect a cognitive function (Garber and Goldin-Meadow, 2002; Pine et al., 2004) or manifest as a spontaneous production of gestures when solving tasks that involve the use of spatial information (Chu and Kita, 2008). The analysis also looked for private speech. Private speech emerges during preschool years and becomes critical to the development of self-regulation. It is an intermediate step between self-regulatory external speech, and internal speech (Savina, 2014); and although private speech is spoken aloud, it is used for self-guidance, planning, and problem-solving, rather than for a communicative purpose (Vygotsky, 1997; Lee et al., 2014).

The coding was based on several indicators, which encompassed various indications. For each inquiry indicator, the coding scheme specifies the behaviors and responses that manifest the specific capabilities (**Supplementary Data Sheet 1**). We then converted the codes into numeric values (**Supplementary Data Sheet 1**).

Three preschool educators and three early STEM education researchers determined the validity of the two scientific exploration tasks and the coding scheme. They examined the tasks and rubrics as per the objectives of the study and age-appropriate requirements. Following the first validation, there was disagreement regarding some indicators and statements. For example, the *planning* indicator is coded as part of the inquiry capabilities and also as a self-regulation capability (**Supplementary Data Sheet 1**). Planning is a very important

stage before engaging in any task or action and has many ways to be realized. Therefore, planning should be perceived as part of the scientific inquiry process and also when regulating cognition in this context and other contexts. All validators agreed that *planning* should be scored both in terms of inquiry skills and self-regulation. Disagreements were resolved by discussion until a consensus was reached, and only statements achieving full agreement were included in the analyses.

Three raters, specialists in early childhood science education research, coded 10% of the video data gathered in this study. The three raters watched the videos carefully (each video was viewed at least twice) and analyzed each video according to the coding scheme (**Supplementary Data Sheet 1**). Inter-rater reliability between the three raters was calculated, producing a Cronbach's Alpha score ranging between 0.70 and 1 for the different indicators.

A fine-grained, video micro-analysis was conducted for each video (for a total of about 100 h of video recordings). Each video was carefully viewed, and every second was coded. We coded every statement the children made, as well as every gesture or facial expression.

Quantitative data were analyzed using the IBM SPSS Statistics software, version 25.0. Using this software, we calculated frequencies, mean values, *t*-tests for independent surveys, ANOVA, Bonferroni analysis, etc. Qualitative data included participants' verbal responses, which were analyzed deductively, according to the indicators measured and the research questions. A mapping analysis was performed to identify relationships between the categories while reassigning them to the appropriate groups and arranging them by various indicators and on various levels (Strauss and Corbin, 1990; Shkedi, 2003). Finally, each indication was counted and added to the quantitative analysis.

RESULTS

The dependent variables in the current study were divided into nominal variables, ordinal variables, and variables on a numerical scale. Prior to examining the study questions and hypotheses, we examined whether the dependent variables measured numerically were normally distributed, using the Shapiro-Wilk tests. Due to the large variability in children's ability at this age, some of the dependent variables were not normally distributed. Therefore, we examined the study questions and hypotheses by conducting parametric and non-parametric tests. The non-parametric analyses findings matched the findings of the parametric analyses. Therefore, we have presented the findings of the parametric analyses, which were measured on numerical scales.

The Children's Intuitive Inquiry Capabilities in an Open-Ended Task

The children's inquiry capabilities were measured by five indicators: their ability to ask questions (total no. of questions), to use tools, to plan, to hypothesize, and to draw conclusions (see **Supplementary Data Sheet 1**). The means, standard deviations (SD), and the Min and Max values of their inquiry capabilities measures are presented in **Table 1**.

As **Table 1** indicates, there was high variability in the children's inquiry capabilities. The highest variability was in the measure of the number of questions asked by the children during the entire scientific exploration. While ∼70% of the children asked between two and seven questions during the entire scientific exploration (146 children), a handful (10; i.e., 4.7%) asked over 10 questions. Therefore, the mean is 4.87 and the standard deviation is 2.96. The questions the participants asked were mostly of two types. The first type included questions related to visual aspects, such as: "What color is that?" or "How come it didn't disappear?" The second type included questions related to activity, or instructions, such as: "Can I mix this?" or "Can I touch it?" We found that despite the children's young age, and the fact that they did not have direct instruction or mediation on the part of the researcher, a few children succeeded in asking questions that testified to a high level of questioning, such as: "I mixed seltzer and oil. If I mix everything together, what will I get?"

Regarding the children's ability to use tools, the mean was in the mid-range. Namely, only 13 children (13.0%) used all four tools competently. Regarding the children's ability to plan and to hypothesize, none of the children attained the maximum score, and their abilities were average. Regarding the children's ability to draw conclusions, their ability was in the mid-range. Only four children (1.9%) scored the maximum value of seven, available in this measure. However, this capability was manifested to the highest extent compared to competent tool use, planning, and hypothesizing.

The Children's Metacognitive Strategic Awareness and Self-Regulation Capabilities in an Open-Ended Task

The children's metacognitive strategic awareness and self-regulation capabilities were measured by means of seven indicators: the level of the children's strategic awareness, self-regulation, their lack of self-regulation, and their regulation of motivation: attention, persistence, autonomy, and engagement

TABLE 1 | Mean, SD, Min, and Max values of the children's inquiry capabilities measure in an open-ended task (*N* = 215).

Children's inquiry capabilities (range)	*M*	*SD*	*Min*	*Max*
Total no. of questions (number)	4.87	2.96	0.00	14.00
Competent use of tools (0–4)	2.07	1.16	0.00	4.00
Planning (0–7)	3.04	1.40	0.00	6.00
Hypothesizing (0–6)	2.14	1.39	0.00	5.00
Drawing conclusions (0–7)	4.32	1.81	0.00	7.00

level) (see **Supplementary Data Sheet 1**). **Table 2** presents the means, standard deviations, Min, and Max values of the children's metacognitive strategic awareness and self-regulation capabilities measures.

As **Table 2** shows, the children's scores on the strategic awareness, attention, persistence, autonomy, and engagement measures were high, with a mean score of over 2.8 out of a maximum score of 4. Moreover, their scores on the lack of self-regulation measure were low, with a mean score of 1.75 out of a maximum score of 7. None of the participants scored the maximum score of 7 in the self-regulation measure, and only six children scored 6 on this measure (2.8%).

Correlations Between the Children's Inquiry Capabilities, Their Metacognitive Strategic Awareness, and Self-Regulation Capabilities in an Open-Ended Task

In order to examine whether correlations would be found between the children's inquiry capabilities and their metacognitive strategic awareness, and self-regulation capabilities, Pearson and Spearman correlations analyses were conducted: Pearson correlations coefficients were calculated for the numerical scale variables, and Spearman correlations coefficients were calculated for the ordinal scale variables.

The analyses and calculation revealed significant positive correlations between the children's five inquiry capabilities measures and the self-regulation capabilities measures. Specifically, we found the highest correlations between the children's inquiry capabilities and their scores on the metacognitive strategic awareness, self-regulation, persistence, autonomy, and engagement measures. These results indicated that as the children's scores on the metacognitive strategic awareness, self-regulation, persistence, autonomy, and engagement measures increased, their scores on the inquiry capabilities measures increased accordingly.

Note that although significant correlations were found among the 215 children who participated in the current study, their strengths were of a medium degree and below (all correlation coefficients were below 0.35). Moreover, no significant negative correlations were found between the five inquiry capabilities measures and their level of lack of self-regulation (see **Table 3**).

Partial correlation analyses were conducted controlling for the general intelligence measured by the *Raven's Progressive Matrices,* receptive vocabulary measured by the *Peabody Picture Vocabulary Test* and the age of the participants. The significant correlations remained significant and in some cases were even strengthened (see **Supplementary Table 1**).

Correlations Between the Children's Inquiry Capabilities and Their Metacognitive Strategic Awareness and Self-Regulation Capabilities in the Structured Task

In order to examine whether correlations would be found between the children's inquiry capabilities and their metacognitive strategic awareness and self-regulation capabilities while conducting the structured task, Pearson and Spearman correlations analyses were conducted: Pearson correlations coefficients were calculated for the numerical scale variables, and Spearman correlations coefficients were calculated for the ordinal scale variables (**Table 4**).

The analyses and calculation revealed significant positive correlations between four of the five children's inquiry capabilities measures and the self-regulation capabilities measures. Specifically, we found correlations between the total number of questions asked, competent use of tools, hypothesizing, and drawing conclusions and their scores on self-regulation. Negative significant correlations were found with the variables of attention (with competent use of tools), persistence, and autonomy (with total number of questions asked). These findings will be further elaborated in the "Discussion" section.

Partial correlation analyses were conducted controlling for the general intelligence measured by the *Raven's Progressive Matrices,* receptive vocabulary measured by the *Peabody Picture Vocabulary Test* and the age of the participants. The significant correlations remained significant and in some cases were even strengthened (see **Supplementary Table 2**).

Differences in the Children's Intuitive Inquiry, Metacognitive Strategic Awareness, and Self-Regulation Capabilities by Type of Task

In the current study, 147 children engaged in a structured task in addition to the open-ended task. In order to examine the differences between the two types of tasks in the dependent variables, we performed three different analyses, depending upon the variable scale. To examine the differences between the two tasks on the ordinal scale variables, we performed Wilcoxon analyses; and to examine the differences between the two tasks on the numerical scale variables, we performed paired samples *t*-test analyses.

Differences in the Children's Inquiry Capabilities by Type of Task

We found significant differences between the two task types in the number of questions the children asked during the entire scientific exploration process, their ability to use tools, and their ability to plan. **Table 4** presents the means and SD of the children's *inquiry capabilities* measures by type of task.

As **Table 5** shows, the number of questions asked by the children during the scientific exploration tasks and their ability to plan were significantly higher in the structured task than in the open-ended task. However, the children's ability to use tools competently was significantly higher in the open-ended task than in the structured task.

Differences in the Children's Metacognitive Strategic Awareness and Self-Regulation Capabilities by Type of Task

Significant differences were found between the two task types in the scores on self-regulation, lack of self-regulation, attention, persistence, and autonomy.

TABLE 2 | Mean, SD, Mdn, and range of the children's metacognitive strategic awareness and self-regulation capabilities in an open-ended task (N = 215).

Children's metacognitive and self-regulation capabilities (range)	*M*	*SD*	*Mdn*	*Min*	*Max*
Strategic awareness (0–4)	3.15	0.80	–	0.00	4.00
Self-regulation (0-sum of points)	10.47	3.11	–	2.00	19.00
Lack of self-regulation (0–7)	1.75	1.59	–	0.00	6.00
Attention[!] (1–4)	3.53	0.61	4.00	2.00	4.00
Persistence[!] (1–4)	3.15	0.75	3.00	1.00	4.00
Autonomy[!] (1–4)	2.88	0.62	3.00	1.00	4.00
Engagement[!] (1–4)	3.35	0.63	3.00	1.00	4.00

[!]*Variables are on ordinal scales; Means as well as Mdn are reported.*

TABLE 3 | Pearson and Spearman correlation coefficients between the children's self-regulation capabilities and their ability to ask questions, to use scientific tools, to plan, to hypothesize, and to draw conclusions in the open-ended task (N = 215).

	Self-regulation capabilities						
Inquiry capabilities	**Strategy awareness**	**Self-regulation**	**Lack of self-regulation**	**Attention[!]**	**Persistence[!]**	**Autonomy[!]**	**Engagement**
Total number of questions	0.16*	0.24***	0.06	−0.10	0.11	−0.05	0.22***
Competent use of tools	0.32***	0.23*	−0.06	0.03	0.20**	0.14*	0.21**
Planning	0.20**	0.01	0.09	0.08	0.21**	0.13*	0.08
Hypothesizing	0.14*	0.05	0.07	0.11	0.17*	0.17*	0.08
Drawing conclusions	0.33***	0.16*	−0.10	0.15*	0.27***	0.25***	0.15*

*$p < 0.05$, **$p < 01$, ***$p < 0.001$. [!]*Variables are on ordinal scales; Spearman instead of Pearson correlation coefficients are reported.*

TABLE 4 | Pearson and Spearman correlation coefficients between the children's self-regulation capabilities and their ability to ask questions, to use scientific tools, to plan, to hypothesize, and to draw conclusions in the structured task (N = 147).

	Self-regulation capabilities						
Inquiry capabilities	**Strategy awareness**	**Self-regulation**	**Lack of self-regulation**	**Attention[!]**	**Persistence[!]**	**Autonomy[!]**	**Engagement**
Total number of questions	0.10	0.32***	0.16	−0.12	−0.33***	−0.38***	0.20*
Competent use of tools	0.38***	0	0.30***	−0.23**	−0.12	−0.15	0.19*
Planning	0.14	0.08	0.14	−0.06	0.10	0.04	−0.04
Hypothesizing	0.12	0.10	0.11	−0.06	0.11	−0.07	0.29***
Drawing conclusions	0.34***	0.20*	0.03	0.05	−0.01	−0.04	0.08

*$p < 0.05$, **$p < 01$, ***$p < 0.001$. [!]*Variables are on ordinal scales; Spearman instead of Pearson correlation coefficients are reported.*

TABLE 5 | Mean, SD, and t-values of the scores of the children's inquiry capabilities measures by type of task (n = 147).

	Open-ended		Structured				
Inquiry capabilities measures	***M***	***SD***	***M***	***SD***	***t***	***P***	**Cohen's *d***
Total number of questions (number)	5.04	3.00	5.78	4.62	−2.16*	0.032	0.18
Competent use of tools (0–4)	2.20	1.20	1.57	1.20	5.56***	0.000	0.46
Planning (0–7)	3.01	1.46	3.83	1.48	−5.25***	0.000	0.43
Hypothesizing (0–6)	1.94	1.48	2.01	1.49	−0.47	0.640	0.04
Drawing conclusions (0–7)	4.14	1.91	4.37	1.53	−1.27	0.207	0.10

*$p < 0.05$; ***$p < 0.001$.

Table 6 presents the means and SD of the children's *metacognitive strategic awareness and self-regulation* measures by task type.

As **Table 6** shows, the scores on the self-regulation and lack of self-regulation measures were significantly higher in the open-ended exploration task as compared to the structured exploration task. During the open-ended task, participants showed significantly higher frequency and level of evidences-manifesting checking and monitoring, controlling, and evaluating (**Supplementary Data Sheet 1**). The scores on the attention, persistence, and autonomy measures were significantly higher in the structured task than in the open-ended task. These differences will be elaborated in the "Discussion" section.

TABLE 6 | Mean, SD, *t*- and *z*-values of the children's metacognitive strategic awareness and self-regulation capabilities by task type (N = 147).

	Open-ended			Structured				
Children's self-regulation capabilities	***M***	***SD***	***Mdn***	***M***	***SD***	***Mdn***	***t/Z***	***p***
Strategic awareness	3.15	0.88	–	3.06	0.72	–	1.12	0.265
Self-regulation	10.77	2.84	–	7.99	3.14	–	9.54***	0.000
Lack of self-regulation	1.80	1.66	–	1.33	1.63	–	2.99**	0.003
Attention[!]	3.46	0.63	4.00	3.63	0.60	4.00	2.98**	0.003
Persistence[!]	3.14	0.79	3.00	3.51	0.64	4.00	4.70***	0.000
Autonomy[!]	2.86	0.70	3.00	3.18	0.65	3.00	4.72***	0.000
Engagement[!]	3.33	0.66	3.00	3.37	0.49	3.00	0.86	0.391

$^{**}p < 0.01$; $^{***}p < 0.001$. [!]Variables are on ordinal scales; Means and Mdn are reported.

FIGURE 1 | Open-ended and structured scientific exploration tasks.

DISCUSSION AND CONCLUSION

This study sought to examine nascent inquiry, metacognitive strategic awareness, and self-regulation capabilities among preschoolers during play-based scientific exploration tasks (open-ended, and structured). In addition, it sought to examine the relationship between inquiry and self-regulation capabilities while preschoolers engage in play-based scientific exploration tasks.

Inquiry Capabilities During the Open-Ended Exploration Task

The study findings show that given the opportunity, children quite naturally exhibit inquiry capabilities during situated scientific exploration: they ask questions, plan, hypothesize, use tools, draw conclusions, and can explain their conclusions. This study's findings are in line with findings of other studies indicating that young children hypothesize and can modify hypotheses when necessary (Legare, 2012); pose scientific queries; comprehend basic scientific concepts (Patrick et al., 2009); and use hypotheses to predict results, evaluate evidence, and produce explanations (Piekny et al., 2014; Walker et al., 2014). Moreover, children's free play involves intuitive experiments that enable them to activate complex mechanisms and comprehend phenomena better (Gopnik, 2012). Most of the studies so far have not tested these capabilities in young children in the context of ecologically situated scientific explorations. The current study is innovative in its focus on children's inquiry capabilities in contexts of authentic, play-based scientific exploration. This study reveals that when presented with materials and equipment of a scientific nature, even without setting explicit goals and rules, children do exhibit scientific inquiry capabilities.

The results suggest that the children's ability to draw conclusions was manifested to a greater extent than other inquiry capabilities tested, given the number of children who succeeded in drawing conclusions without mediation, as well as the level of their conclusions. Drawing conclusions is the final stage of the inquiry process, leading to the production of new information and posing new research questions. ErgaZaki and ZogZa (2013) reason that involvement in inquiry improves inductive ability. The findings that have emerged in our study are in line with that claim. The children in our study referred to what happened to the materials during their explorative actions. For example, when they mixed water with oil, they could see that the oil floated on the water, but when they mixed juice with milk, the juice tinted the milk and changed its color. Therefore, in accordance with the concrete, visible results, the children were able to draw conclusions. Other inquiry capabilities, such as planning and hypothesizing, involve more complex, abstract thinking skills that are of a predictive type, rather than being based on concrete, visual information. Thus, capabilities such as planning and hypothesizing are more challenging, as they are based on abstract thought and thinking a few steps ahead.

Scholarly consensus holds that the skill of posing questions is fundamental to inquiry processes and contributes to cognitive development and higher-order thinking (National Research Council (NRC), 2012). In terms of the nature of science, it is important that children understand that with the aid of questions, they access new information they did not previously

have (Kuhn and Dean, 2005). In our study, the children's ability to pose questions developed during the course of the open-ended exploration task. The participants actively explored the use of tools and materials, and observed what happened to the materials after mixing them, which generated more questions about the task, such as how the tools work, etc.

Engaging in science is a complex process that leads to creating new knowledge. Zimmerman and Croker (2013) contend that knowledge of science and mastery of inquiry skills are mutually enriching and lead to the development of scientific thinking. This bears out the importance of providing children with play-based scientific exploration experiences that include exposure to a range of phenomena, tools, and materials that, with time, become familiar, enrich their knowledge of science, and raise the level of their inquiry capabilities (Tunnicliffe, 2015).

Metacognitive Strategic Awareness and Self-Regulation Capabilities During the Open-Ended Exploration Task

The second research question sought to study nascent metacognitive strategic awareness and self-regulation capabilities among children during scientific exploration tasks. The study examined four parameters (comprising 13 indicators; **Supplementary Data Sheet 1**) that characterize the emerging metacognitive strategic awareness and self-regulation capabilities in children. According to our findings, the young participants exhibited capabilities of metacognitive strategic awareness, self-regulation, and a very high level of attention, persistence on task, and engagement. They exhibited abilities of monitoring and control, such as defining difficulty and then addressing a solution; they exhibited capabilities of testing and evaluating; they expressed satisfaction or dissatisfaction with the results and could explain why.

This study corroborates the results of other studies, indicating that through the use of age-appropriate methodology, such as real-time (online) data gathering while young children are engaged with a task, we can see the nascent abilities of planning, monitoring, control, and reflection in young children (Whitebread and Coltman, 2010; Whitebread et al., 2010; Bryce and Whitebread, 2012). Our results are in line with studies showing that preschool-aged children can plan, set goals, and conduct reflective processes on their learning (Whitebread et al., 2007; Shamir et al., 2009; Larkin, 2010; Whitebread and Coltman, 2010).

The literature indicates that children activate strategies when their task is appropriate and of significance for them (Whitebread and Coltman, 2010). In our study, despite the children's young age, and the fact that most of them experienced scientific exploration of this type for the first time, one conspicuous result was their low score on the measure of lack of self-regulation. The average lack of self-regulation score was 1.8 (out of a maximum of 7). Consequently, the researchers could only observe indications such as brute force and repeated errors in rare instances.

Self-regulation includes the ability to control the intensity of one's responses. Metacognitive experiences relate to the emotional aspect of learning, including motivation and social-emotional processes (Efklides, 2006). These occur in the learner through cognitive experiences, including feelings, judgments, and knowledge about the task. Self-regulation is therefore one of the manifestations of reciprocity between emotion and cognition (Lee, 2005). In this study, we examined four measures related to emotion and motivation: attention, persistence, autonomy, and engagement with the task.

Most of the children worked happily while maintaining their focus on the task. They persevered independently while facing the difficulties that arose during the task, and exhibited involvement, interest, and enjoyment during its performance. When children engage in age- and developmental stage-appropriate activity, they perform better in aspects like self-regulation and cognitive procedures, as well as emotionally and behaviorally (Efklides, 2006).

Consistent with the literature (e.g., Bronson, 2000; Winne and Perry, 2000; Winne, 2018), this study shows that even without formal instruction in the subject area, without explicit task instructions, and without explicit goals, preschoolers succeeded in implementing components of scientific inquiry and exhibiting manifestations of metacognitive strategic awareness and self-regulation during an open-ended scientific exploration task that included the use of tools and materials.

Relationship Between Inquiry, Metacognitive Strategic Awareness, and Self-Regulation Capabilities During an Open-Ended Exploration Task

The study results support our initial hypothesis and indicate significant correlations between the five measures of preschoolers' inquiry capabilities and manifestations of metacognitive strategic awareness and self-regulation. Specifically, we found correlations between the children's inquiry capabilities and most of their scores on the strategic awareness, self-regulation, persistence, and engagement measures in the open-ended play-based task. These results indicated that as the children's scores on the inquiry capabilities measures increased, the scores on the metacognitive strategic awareness, self-regulation, persistence, and engagement measures increased accordingly. As the children's strategic awareness and self-regulation grew while they were performing the task, skills such as planning the process, gathering data, making the connection between cause and effect, and drawing conclusions improved. Through using these skills, they succeeded in addressing the questions themselves: "What is the nature of this task?," "What type of strategies do I need to adopt in order to succeed?," "Did I choose the right strategy or shall I change it?," and so forth.

Data analysis revealed no significant correlations between the five inquiry capabilities measures and the children's level of lack of self-regulation. As previously mentioned, the result shows that even without formal instruction of scientific inquiry skills, preschoolers succeeded in monitoring and controlling their actions and responses, while rarely showing brut or negative responses, nor repeated errors.

Note that among the 215 children who participated in the current study, the correlations between the five inquiry capabilities measures and manifestations of metacognitive

strategic awareness and self-regulation, were significant but weak. These results might be explained by virtue of the very young age of the participants of this study, whose thinking and learning capabilities were still in the early stages of development. The variability in their inquiry level of performances (SD presented in **Table 1**) might explain the weak correlations between the measures tested. The significant correlations, although weak, support our hypothesis, yet further studies are needed with a possible refinement of the coding scheme and of the play-based exploration task itself.

This study corroborates studies that attest to significant correlations between inquiry skills and self-regulation and metacognitive thinking. The defining feature of scientific thinking is the set of cognitive and metacognitive skills involved in differentiating and coordinating theory and evidence. In particular, it is precisely metacognitive awareness that differentiates more from less sophisticated scientific thinking (Kuhn, 2002). Scientific thinking is an umbrella term that encompasses the reasoning and problem-solving skills used in generating, testing, and revising hypotheses or theories, and in the case of fully developed skills, reflecting on the process of knowledge acquisition and modification. Acquired inquiry skills lead to knowledge modification, and in turn, this developing knowledge influences the development of more sophisticated inquiry skills. Concurrently, as metacognitive capabilities develop, children and adolescents gain a better understanding of the nature of inquiry and the use of skills (Zimmerman, 2007). For these reasons, examining the relations between inquiry, metacognitive strategic awareness, and self-regulation capabilities during an open-ended, play-based exploration task, bear great importance toward developing appropriate early science pedagogy.

Inquiry, Metacognitive Strategic Awareness, and Self-Regulation Capabilities During Structured Scientific Exploration Task

This study has attempted to examine nascent inquiry, metacognitive strategic awareness, and self-regulation capabilities in preschoolers during a structured exploration task and compare them to the same abilities exhibited during an open-ended exploration task. Our motivation was to examine whether preschoolers' capabilities will differ in the context of structured task which is aligned with the view that young children need guidance and explicit instruction to acquire the components of scientific inquiry (Zimmerman, 2007) compared to the context of open-ended, play-based task–allowing the children to apply and test their intuitive theories and improve their nascent inquiry capabilities (Buchsbaum et al., 2011).

The study findings indicate significant differences between the two task types regarding the children's capabilities to plan, use tools, and pose questions. While each of the scientific tasks – structured and open-ended – both invite and emphasize inquiry capabilities at various intensities among the study participants, the ability to plan was better manifested during the structured task.

It is likely that when the task establishes a final goal, it is easier for children to plan their steps toward reaching it. The ability to use tools was better manifested in the open-ended, play-based exploration task. The open-ended task provided the participants with endless options to make attempts and to carry out the experiments with the tools and materials that were situated in front of them. In the structured exploration task, however, the children were focused on the instructions and the actions required for carrying out the task, and they did not necessarily need to undertake many actions or use many tools in order to do so. The number of questions asked during the exploration process was greater in the structured task; however, during both tasks, questions arose regarding the use of tools and materials and about how to carry out the task. This finding is in line with Samarapungavan et al. (2008) who affirmed that questioning encourages reflective scientific thinking and learning.

Our findings show that although the children manifested higher levels of attention, persistence, and autonomy during the structured task, they manifested lower levels of self-regulation. These findings may indicate that the young participants could better regulate their emotions (while, to a lesser degree, regulating their cognition) when they had clear task instructions. Considering that the cognitive and emotional aspects of self-regulation can be correlated (Efklides, 2006), these findings may appear to be contradictory; however, they are not: The more demanding task of the two was the structured exploration task, which involved specific instructions and required more self-regulation than the open-ended task. Despite the particularly young age of the participants, the clear instructions provided for the structured exploration task reduced their emotional load, thus enabling them to muster the effort needed to accomplish the task.

However, during both tasks, the children were active during the exploration. This likely stems from the fact that the open-ended task enabled the children to make their own decisions on the matter of what to investigate with the items at hand, while the structured task included clear instructions with required, defined results.

Data analysis revealed weak but significant correlation coefficients between the children's self-regulation capabilities and some inquiry capabilities in the structured task and in the open-ended task. These significant correlations remained after controlling for age, PPVT, and Raven scores. These findings show that both crystalized and fluid intelligence positively support regulation of cognition and actual scientific exploration capabilities. Intelligence is associated with a greater ability to regulate one's impulses, emotions, and behavior, and may further explain why these two traits (intelligence and self-control) are generally so closely related to important life outcomes such as success in primary and secondary education. A growing body of convincing evidence suggests that intelligence is closely tied to the development of self-regulation in the early stages of the course of life (Primi et al., 2010).

The literature addresses the effects of structured environments vs. those of open environments on engagement in science and exploratory processes among children (Chinn and Malhotra, 2002; Bergen, 2009). Most studies recommend providing a structured environment, based on research showing that without

guidance and explicit instructions, children find it difficult to acquire the components of scientific thinking, as well as to implement scientific thinking skills such as planning experiments, documentation, and processing existing knowledge with new knowledge. In other words, they should be taught these processes in a structured manner (Zimmerman, 2007). According to this view, a curriculum that emphasizes a structured planning of science engagement – including details regarding the type of activity, the content, and the sequence – mitigates the cognitive burden and maintains the focus on the goals. Such an approach even improves children's cognitive, lingual, and emotional level of engagement in science (Weisberg et al., 2013). The results of the present study do support this approach. However, although the participants in this study showed better inquiry capabilities in the structured scientific exploration task, this was not the case regarding self-regulation capabilities. Significant higher scores of self-regulation manifestations were revealed in the context of the open-ended, play-based, exploration task. Other researchers also support open-ended exploration, based on studies attesting to the effectiveness of learning that stems from spontaneous investigation, as well as testimonies that this method improves the young students' cognitive, emotional, and social abilities (Bergen, 2002; Hyson, 2003), as well as their creative imagination. The results of the present study suggest the importance of combining various learning environments and experiences in science instruction that encourage children to engage in structured exploration alongside spontaneous, play-based, open-ended, exploration.

Limitations of the Study

Working with 5- to 6-year-old children requires adapting the research tools to meet the cognitive, linguistic, emotional, and behavioral capabilities of children at this age. Thus, due to the age limitations of the research sample, the findings are based primarily on deriving meaning from observing the children's behaviors, rather than being solely based upon what a child reported verbally. Although the validation process of the tasks and coding scheme was accomplished with six experts, and the inter-rater reliability was accomplished with three experts who coded 10% of the video data, for the sake of reliability, future studies should analyze a greater percentage of the video data, given the relatively high inference used in the coding system.

The exploration tasks and coding scheme of this study were developed especially for preschool-age children and were based on the children's behavioral and verbal responses. However, the tasks lack direct questions addressed to the children, seeking to find out what they think about structured vs. open-ended exploration tasks. Future research should address this gap.

Regarding the study procedure, although the open-ended and structured tasks were presented in a counterbalanced manner, one may argue that those children who began with the structured task may have experienced behaviors associated with structure and clear goals, leading to higher incidences of inquiry skills, metacognitive strategic awareness, and self-regulation. In turn, this may have affected their performance of the open-ended task. This issue can be investigated by further data analysis. Findings may have potential pedagogical implications and can enrich the discourse on children's intuitive capabilities and the ways they develop.

In this study, the children's experiences during the open-ended exploration task enabled them to decide what to explore using the tools and materials at their disposal. While they decided how to act, they were limited in terms of location; in other words, they worked at a table, over a tray, and could not choose to work elsewhere. Future research should examine these abilities through an open-ended task that allows the children to decide what, how, and where (for example, outside in the schoolyard) to explore. This would offer researchers the opportunity to provide the children with additional phenomena to investigate.

Another limitation of this study may lie with the measurements of each variable in the scoring scheme. For example, some behavioral items in the scoring scheme of scientific inquiry skills overlapped with those of metacognitive strategic awareness and self-regulation. Specifically, *planning* appears in the metacognitive variables and also as a variable of the inquiry capabilities. Each *planning* variable refers to a different issue: in one instance it is the planning of physical actions, while the other refers to planning on a mental level (e.g., how to monitor my strategies).

We considered this potential bias when developing the coding scheme. For that reason, we conducted an intensive validation process, with six preschool educators and early STEM education researchers. However, more accuracy is needed to differentiate between these variables. These overlaps pose a potential risk of overestimating correlations among the variables and deviating from the study assumptions. Thus, further studies are needed to refine these measures.

SUMMARY

To conclude, this study offers insights related to various aspects of nascent scientific thinking in children, focusing on inquiry, metacognitive strategic awareness, and self-regulation capabilities. Few studies have examined these nascent abilities in children while performing situated, scientific exploration of materials. This study shows that preschoolers like to explore and engage in science activities, and can maintain their attention, persistence, and engagement. Preschoolers exhibit verbal and behavioral responses and actions that demonstrate nascent inquiry capabilities, as well as indications of metacognitive strategic awareness and self-regulation capabilities. Some of these capabilities are better manifested during structured exploration, and others during open-ended exploration tasks. Significant correlations were found between these capabilities. These findings indicate the importance of offering children a learning environment that provides them with rich opportunities to explore and develop their intuitive inquiry capabilities: An environment that draws children into active involvement in the inquiry process and development of nascent metacognitive awareness and self-regulation capabilities.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Rene Osizon, Ethics Committee (#10495), The Office of the Chief Scientist of the Ministry of Education. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

All authors contributed by editing the literature review, collecting the data, analyzing the data, drawing the conclusions, and writing and editing the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01790/full#supplementary-material

REFERENCES

Alexander, J. M., Johnson, K. E., and Kelley, K. (2012). Longitudinal analysis of the relations between opportunities to learn about science and the development of interests related to science. *Sci. Educ.* 96, 763–786. doi: 10.1002/sce.21018

Andiema, N. C. (2016). Effect of child centred methods on teaching and learning of science activities in pre-schools in Kenya. *J. Educ. Pract.* 7, 1–9.

Basilio, M., and Rodríguez, C. (2011). Usos, gestos y vocalizaciones privadas: de la interacción social a la autorregulación [Private uses, gestures and vocalizations: from social interaction to self-regulation]. *Infancia y Aprendizaje* 34, 181–194. doi: 10.1174/021037011795377593

Beishuizen, J. J., and Veenman, M. V. (2004). Intellectual and met cognitive skills of novices while studying texts under conditions of text difficulty and time constraint. *Learn. Instruct.* 6, 621–640. doi: 10.1016/j.learninstruc.2004.09.004

Bergen, D. (2002). The role of pretend play in children's cognitive development. *Early Child. Res. Pract.* 4. Available online at: https://files.eric.ed.gov/fulltext/ED464763.pdf (accessed September 10, 2019).

Bergen, D. (2009). Play as the learning medium for future scientists, mathematicians, and engineers. *Am. J. Play* 1, 413–428.

Best, J. R., and Miller, P. H. (2010). A developmental perspective on executive function. *Child Dev.* 81, 1641–1660. doi: 10.1111/j.1467-8624.2010.01499.x

Blair, C., and Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy abilities in kindergarten. *Child Dev.* 78, 647–663. doi: 10.1111/j.1467-8624.2007.01019.x

Bronson, M. (2000). *Self-Regulation in Early Childhood: Nature and Nurture*. New York, NY: Guilford Press.

Brown, A. (1987). "Metacognition, executive control, self-regulation, and other more mysterious mechanisms," in *Metacognition, Motivation, and Understanding*, eds F. Weinert and R. Kluwe (Hillsdale, NJ: Erlbaum), 65–116.

Bryce, D., and Whitebread, D. (2012). The development of metacognitive skills: Evidence from observational analysis of young children's behavior during problem-solving. *Metacogn. Learn.* 7, 197–217. doi: 10.1007/s11409-012-9091-2

Bryce, D., Whitebread, D., and Szûcs, D. (2015). The relationships among executive functions, metacognitive skills and educational achievement in 5 and 7 year-old children. *Metacogn. Learn.* 10, 181–198. doi: 10.1007/s11409-014-9120-4

Buchsbaum, D., Gopnik, A., Griffiths, T. L., and Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition* 120, 331–340. doi: 10.1016/j.cognition.2010.12.001

Bumbacher, E., Salehi, S., Wieman, C., and Blikstein, P. (2018). Tools for science inquiry learning: tool affordances, experimentation strategies, and conceptual understanding. *J. Sci. Educ. Technol.* 27, 215–235. doi: 10.1007/s10956-017-9719-8

Bustamante, A. S., Greenfield, D. B., and Nayfeld, I. (2018). Early childhood science and engineering: engaging platforms for fostering domain-general learning skills. *Educ. Sci.* 8:144. doi: 10.3390/educsci8030144

Butler, D. L., Lauscher, H. N., Jarvis-Selinger, S., and Beckingham, B. (2004). Collaboration and self-regulation in teachers' professional development. *Teach. Teach. Educ.* 20, 435–455. doi: 10.1016/j.tate.2004.04.003

Butler, D. L., and Schnellert, L. (2012). Collaborative inquiry in teacher professional development. *Teach. Teach. Educ.* 28, 1206–1220. doi: 10.1016/j.tate.2012.07.009

Chinn, C. A., and Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: a theoretical framework for evaluating inquiry tasks. *Sci. Educ.* 86, 175–218. doi: 10.1002/sce.10001

Chu, M., and Kita, S. (2008). Spontaneous gestures during mental rotation tasks: insights into the microdevelopment of the motor strategy. *J. Exp. Psychol.* 137, 706–723. doi: 10.1037/a0013157

De Corte, E., Verschaffel, L., and Op't Eynde, P. (2000). "Self-regulation: a characteristic and a goal of mathematics education," in *Handbook of Self-Regulation*, eds M. Zeidner, M. Boekaerts, and P. Pintrich (Cambridge, MA: Academic Press), 687–726.

Dewey, J. (1938). *How We Think: A Restatement of Reflective Thinking to the Educative Process*. Boston: D. C. Heath.

Diamond, K. E., Justice, L. M., Siegler, R. S., and Snyder, P. A. (2013). *Synthesis of IES Research on Early Intervention and Early Childhood Education (NCSER 2013–3001)*. Washington, DC: National Center for Special Education Research.

Dörr, L., and Perels, F. (2019). Improving metacognitive abilities as an important prerequisite for self-regulated learning in preschool children. *Int. Electron. J. Element. Educ.* 11, 449–459. doi: 10.26822/iejee.2019553341

Dunn, K. E., Rakes, G. C., and Rakes, T. A. (2014). Influence of academic self-regulation, critical thinking, and age on online graduate students' academic help-seeking. *Dist. Educ.* 35, 75–89. doi: 10.1080/01587919.2014.891426

Dunn, L. M., and Dunn, D. M. (2007). *PPVT-4: Peabody Picture Vocabulary Test*. London: Pearson.

Early Childhood Stem Working Group (2017). *Early STEM Matters: Providing High-Quality STEM Experiences for All Young Learners*. Policy Report. Chicago, IL: UChicago STEM Education.

Efklides, A. (2006). Metacognition and affect: what can metacognitive experiences tell us about the learning process? *Educ. Res. Rev.* 1, 3–14. doi: 10.1016/j.edurev.2005.11.001

ErgaZaki, M., and ZogZa, V. (2013). How does the model of inquiry-based science education work in the kindergarten: the case of biology. *Rev. Sci. Math. ICT Educ.* 7, 73–97.

Eshach, H. (2003). Inquiry-events as a tool for changing science teaching efficacy belief of kindergarten and elementary school teachers. *J. Sci. Educ. Technol.* 12, 495–501. doi: 10.1023/b:jost.0000006309.16842.c8

Eshach, H. (2006). *Science Literacy in Primary Schools and Pre-Schools*. Dordrecht: Springer.

Flavell, J. H. (1976). "Metacognitive aspects of problem solving," in *The Nature of Intelligence*, ed. L. B. Resnick (Hillsdale, NJ: Lawrence Erlbaum), 231–235.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: a new area of cognitive–developmental inquiry. *Am. Psychol.* 34:906. doi: 10.1037/0003-066x.34.10.906

Fleer, M. (2009). Supporting scientific conceptual consciousness or learning in 'a roundabout way' in play–based contexts. *Int. J. Sci. Educ.* 31, 1069–1089. doi: 10.1080/09500690801953161

French, L. (2004). Science as the center of a coherent, integrated early childhood curriculum. *Early Child. Res. Q.* 19, 138–149. doi: 10.1016/j.ecresq.2004.01.004

Garber, P., and Goldin-Meadow, S. (2002). Gesture offers insight into problem-solving in adults and children. *Cogn. Sci.* 26, 817–831. doi: 10.1207/s15516709cog2606_5

Gopnik, A. (2012). Scientific thinking in young children: theoretical advances, empirical research, and policy implications. *Science* 337, 1623–1627. doi: 10.1126/science.1223416

Greeno, J. G. (2002). *Students With Competence, and Accountability: Affording Intellective Identities un Classrooms*. New York, NY: Collage Board.

Hattie, J. A. (2009). *Visible Learning: A Synthesis of 800+ Meta-Analyses on Achievement*. Abingdon: Routledge.

Hendy, L., and Whitebread, D. (2000). Interpretations of independent learning in the early years. *Int. J. Early Years Educ.* 8, 243–252. doi: 10.1080/09669760050156785

Hidi, S., and Ainley, M. (2008). "Interest and self-regulation: relationships between two variables that influence learning," in *Motivation and Self-Regulated Learning: Theory, Research, and Applications*, eds D. H. Schunk and B. J. Zimmerman (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 77–109.

Hyson, M. (2003). Putting early academics in their place. *Educ. Leadersh.* 60, 20–23.

Istomina, Z. M. (1975). The development of voluntary memory in preschool-age children. *Soviet Psychol.* 13, 5–64. doi: 10.2753/rpo1061-040513045

Jirout, J., and Zimmerman, C. (2015). "Development of science process skills in the early childhood years," in *Research in Early Childhood Science Education*, eds M. Saçkes and K. C. Trundle (Dordrecht: Springer), 143–165. doi: 10.1007/978-94-017-9505-0_7

Johnson, R. B., Onwuegbuzie, A. J., and Turner, L. A. (2007). Toward a definition of mixed methods research. *J. Mix. Methods Res.* 1, 112–133. doi: 10.1177/1558689806298224

Jones, I., Lake, V. E., and Lin, M. (2008). "Early childhood science process skills," in *ContemporaryPerspectives on Scienceandtechnology in Earlychildhoodeducation*, eds O. N. Sarachoand and B. Spodek (Mumbai: IAP), 17–40.

Kearney, M., Treagust, D. F., Yeo, S., and Zadnik, M. G. (2001). Student and teacher perceptions of the use of multimedia supported predict–observe-explain tasks to probe understanding. *Res. Sci. Educ.* 31, 589–615.

Kitsantas, A., Steen, S., and Huie, F. (2017). The role of self-regulated strategies and goal orientation in predicting achievement of elementary school children. *Int. Electron. J. Element. Educ.* 2, 65–81.

Klahr, D., Zimmerman, C., and Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science* 333, 971–975. doi: 10.1126/science.1204528

Klein, P. S., and Yablon, Y. B. (2008). *From Research to Practice in Early Childhood Education*. Jerusalem: The Israel Academy of Sciences and Humanities.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychol. Rev.* 96:674. doi: 10.1037/0033-295x.96.4.674

Kuhn, D. (2002). "What is scientific thinking, and how does it develop?," in *Blackwell Handbooks of Developmental Psychology. Blackwell Handbook of Childhood Cognitive Development*, ed. U. Goswami (Hoboken, NJ: Blackwell Publishing), 371–393. doi: 10.1002/9780470996652.ch17

Kuhn, D., Black, J., Keselman, A., and Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cogn. Instruct.* 18, 495–523. doi: 10.1207/s1532690xci1804_3

Kuhn, D., and Dean, D. Jr. (2005). *Scaffolding the Development of Inquiry Skills in Academically at-Risk Populations*. New York, NY: Columbia University.

Larkin, S. (2010). *Metacognition in Young Children*. London: Routledge.

Lee, E. (2005). The relationship of motivation and flow experience to academic procrastination in university students. *J. Genet. Psychol.* 166, 5–15. doi: 10.3200/gntp.166.1.5-15

Lee, S., McDonough, A., and Bird, J. (2014). Investigating eight- to nine-year-olds' self- regulatory self-talk in the context of their classroom tasks. *Early Child Dev. Care* 184, 1661–1676. doi: 10.1080/03004430.2013.875539

Legare, C. H. (2012). Exploring explanation: explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Dev.* 83, 173–185. doi: 10.1111/j.1467-8624.2011.01691.x

Loyens, S. M., and Rikers, R. M. J. P. (2011). "Instruction based on inquiry," in *Handbook of Research on Learning and Instruction*, eds P. Alexander and R. E. Mayer (London: Routledge), 361–381.

Mantzicopoulos, P., Patrick, H., and Samarapungavan, A. (2013). Science literacy in school and home contexts: kindergarteners' science achievement and motivation. *Cogn. Instruc.* 31, 62–119. doi: 10.1080/07370008.2012.742087

McClure, E., Guernsey, L., Clements, D., Bales, S., Nichols, J., Kendall-Taylor, N., et al. (2017). *STEM Starts Early: Grounding Science, Technology, Engineering, and Math Education in Early Childhood*. New York, NY: The Joan Ganz Cooney Center at Sesame Workshop.

Mevarech, Z. R. (1995). Metacognition, general ability, and mathematical understanding. *Early Educ. Dev.* 6, 155–168. doi: 10.1207/s15566935eed0602_4

Michalsky, T., Zion, M., and Mevarech, Z. R. (2007). Developing students' metacognitive awareness in asynchronous learning networks in comparison to face-to-face discussion groups. *J. Educ. Comput. Res.* 36, 395–424. doi: 10.2190/320v-8h4w-1123-44r6

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., et al. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proc. Natl. Acad Sci. U.S.A.* 108, 2693–2698. doi: 10.1073/pnas.1010076108

Mulder, H., Hoofs, H., Verhagen, J., van der Veen, I., and Leseman, P. P. (2014). Psychometric properties and convergent and predictive validity of an executive function test battery for two-year-olds. *Front. Psychol.* 5:733. doi: 10.3389/fpsyg.2014.00733

National Research Council (NRC) (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National The National Academies Press.

National Science Teachers Association [NSTA] (2014). *NSTA Position Statement: Early Childhood Science Education*. Available online at: http://www.nsta.org/about/positions/earlychildhood.aspx (accessed October 18, 2020).

NGSS Lead States (2013). *Next Generation Science Standards: For States, by States*. Washington, DC: The National Academies Press.

Osborne, J., Simon, S., and Collins, S. (2003). Attitudes towards science: a review of the literature and its implications. *Int. J. Sci. Educ.* 25, 1049–1079. doi: 10.1080/0950069032000032199

Patrick, H., Mantzicopoulos, P., and Samarapungavan, A. (2009). Motivation for learning science in Kindergarten: is there a gender gap and does integrated inquiry and literacy instruction make a difference. *J. Res. Sci. Teach.* 46, 166–191. doi: 10.1002/tea.20276

Piekny, J., Grube, D., and Maehler, C. (2014). The development of experimentation and evidence evaluation skills at preschool age. *Int. J. Sci. Educ.* 36, 334–354. doi: 10.1080/09500693.2013.776192

Piekny, J., and Maehler, C. (2013). Scientific reasoning in early and middle childhood: the development of domain−general evidence evaluation, experimentation, and hypothesis generation skills. *Br. J. Dev. Psychol.* 31, 153–179. doi: 10.1111/j.2044-835x.2012.02082.x

Pine, K. J., Lufkin, N., and Messer, D. (2004). More gestures than answers: children learning about balance. *Dev. Psychol.* 40, 1059–1067. doi: 10.1037/0012-1649.40.6.1059

Primi, R., Ferrão, M. E., and Almeida, L. S. (2010). Fluid intelligence as a predictor of learning: a longitudinal multilevel approach applied to math. *Learn. Indiv. Differ.* 20, 446–451. doi: 10.1016/j.lindif.2010.05.001

Raven, J. C. (1956). *Coloured Progressive Matrices (CPM)*. San Antonio, TX: Pearson Education, Inc.

Rickey, D., and Stacy, A. M. (2000). The role of metacognition in learning chemistry. *J. Chem. Educ.* 77:915. doi: 10.1021/ed077p915

Rimm-Kaufman, S. E., Curby, T. W., Grimm, K. J., Nathanson, L., and Brock, L. L. (2009). The contribution of children's self-regulation and classroom quality to children's adaptive behaviors in the kindergarten classroom. *Dev. Psychol.* 45:958. doi: 10.1037/a0015861

Rodríguez, C., and Palacios, P. (2007). Do private gestures have a self-regulatory function? A case study. *Infant Behav. Dev.* 30, 180–194. doi: 10.1016/j.infbeh.2007.02.010

Samarapungavan, A., Mantzicopoulos, P., and Patrick, H. (2008). Learning science through inquiry in kindergarten. *Sci. Educ.* 92, 868–908. doi: 10.1002/sce.20275

Samarapungavan, A., Patrick, H., and Mantzicopoulos, P. (2011). What kindergarten students learn in inquiry-based science classrooms. *Cogn. Instruct.* 29, 416–470.

Savina, E. (2014). Does play promote self-regulation in children? *Early Child Dev. Care* 184, 1692–1705. doi: 10.1080/03004430.2013.875541

Schraw, G., Crippen, K. J., and Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Res. Sci. Educ.* 36, 111–139. doi: 10.1007/s11165-005-3917-8

Schraw, G., and Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemp. Educ. Psychol.* 19, 460–475. doi: 10.1006/ceps.1994.1033

Schraw, G., and Moshman, D. (1995). Metacognitive theories. *Educ. Psychol. Rev.* 7, 351–371. doi: 10.1007/bf02212307

Schulz, L. (2012). The origins of inquiry: inductive inference and exploration in early childhood. *Trends Cogn. Sci.* 16, 382–389. doi: 10.1016/j.tics.2012.06.004

Schweinhart, L. J., and Weikart, D. P. (1997). *Lasting Differences: The High*. Ypsilanti, MI: High/Scope Press.

Shamir, A., Mevarech, Z. R., and Gida, C. (2009). The assessment of meta-cognition in different contexts: individualized vs. peer assisted learning. *Metacogn. Learn.* 4, 47–61. doi: 10.1007/s11409-008-9032-2

Shapiro, C. A., and Sax, L. J. (2011). Major selection and persistence for women in STEM. *New Dir. Inst. Res.* 152, 5–18. doi: 10.1002/ir.404

Sheridan, S., and Samuelsson, I. P. (2001). Children's conceptions of participation and influence in pre-school: a perspective on pedagogical quality. *Contemp. Issues Early Child.* 2, 169–194. doi: 10.2304/ciec.2001.2.2.4

Shkedi, A. (2003). *Words of Meaning: Qualitative Research-Theory and Practice*. Tel-Aviv: Tel-Aviv university Ramot.

Spektor-Levy, O., Basilio, M., Zachariou, A., and Whitebread, D. (2017). Young children's spontaneous manifestation of self-regulation and metacognition during constructional play tasks. *Teach. Coll. Rec.* 119, 1–28.

Strauss, A., and Corbin, J. (1990). *Basics of Qualitative Research*. Thousand Oaks, CA: SAGE publications.

Sylva, K., Melhuish, E. C., Sammons, P., Siraj-Blatchford, I., and Taggart, B. (2004). *The Effective Provision of Pre- School Education (EPPE) Project: Technical Paper 12 - The Final Report: Effective Pre-School Education*. London: DfES.

Tao, Y., Oliver, M., and Venville, G. (2012). Long-term outcomes of early childhood science education: Insights from a cross-national comparative case study on conceptual understanding of science. *Int. J. Sci. Math. Educ.* 10, 1269–1302. doi: 10.1007/s10763-012-9335-2

Trundle, K. C., and Saçkes, M. (2015). *Research in Early Childhood Science Education*. Berlin: Springer.

Tunnicliffe, S. D. (2015). *Starting Inquiry-Based Science in the Early Years: Look, Talk, Think and Do*. London: Routledge.

Veenman, M. V., Van Hout-Wolters, B. H., and Afflerbach, P. (2006). Metacognition and learning: conceptual and methodological considerations. *Metacogn. Learn.* 1, 3–14. doi: 10.1007/s11409-006-6893-0

Veenman, M. V. J., Wilhelm, P., and Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learn. Instruct.* 14, 89–109. doi: 10.1016/j.learninstruc.2003.10.004

Vygotsky, L. S. (1978). *Mind in Society*. Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1986). *Thought and Language*. Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1997). *The Collected Works of LS Vygotsky: Problems of the Theory and History of Psychology*, Vol. 3. New York, NY: Plenum Press.

Wager, A., and Parks, A. N. (2014). "Learning mathematics through play," in *Handbook of Play and Learning in Early Childhood*, eds E. Brooker, M. Blaise, and S. Edwards (Thousand Oaks, CA: SAGE publications), 216–227. doi: 10.4135/9781473907850.n21

Walker, C. M., Lombrozo, T., Legare, C. H., and Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition* 133, 343–357. doi: 10.1016/j.cognition.2014.07.008

Weisberg, D. S., Hirsh-Pasek, K., and Golinkoff, R. M. (2013). Embracing complexity: rethinking the relation between play and learning: comment on Lillard et al.(2013). *Psychol. Bull.* 139, 35–39. doi: 10.1037/a0030077

Whitebread, D., Almeqdad, Q., Bryce, D., Demetriou, D., Grau, V., and Sangster, C. (2010). "Metacognition in young children: Current methodological and theoretical developments," in *Trends and Prospects in Metacognition Research*, eds A. Efklides and P. Misailidi (Boston, MA: Springer), 233–258.

Whitebread, D., Anderson, H., Coltman, P., Page, C., Pasternak, D. P., and Mehta, S. (2005). Developing independent learning in the early years. *Education* 33, 40–50. doi: 10.1080/03004270585200081

Whitebread, D., Bingham, S., Grau, V., Pasternak, D. P., and Sangster, C. (2007). Development of metacognition and self-regulated learning in young children: role of collaborative and peer-assisted learning. *J. Cogn. Educ. Psychol.* 6, 433–455. doi: 10.1891/194589507787382043

Whitebread, D., and Coltman, P. (2010). Aspects of pedagogy supporting metacognition and self-regulation in mathematical learning of young children: evidence from an observational study. *ZDM* 42, 163–178. doi: 10.1007/s11858-009-0233-1

Whitebread, D., Coltman, P., Pasternak, D. P., Sangster, C., Grau, V., Bingham, S., et al. (2009). The development of two observational tools for assessing metacognition and self-regulated learning in young children. *Metacogn. Learn.* 4, 63–85. doi: 10.1007/s11409-008-9033-1

Williams, J. J., and Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cogn. Psychol.* 66, 55–84. doi: 10.1016/j.cogpsych.2012.09.002

Winne, P. H. (2018). Theorizing and researching levels of processing in self−regulated learning. *Br. J. Educ. Psychol.* 88, 9–20. doi: 10.1111/bjep.12173

Winne, P. H., and Hadwin, A. F. (2008). The weave of motivation and self-regulated learning. *Motiv. Self Regul. Learn.* 2, 297–314.

Winne, P. H., and Perry, N. E. (2000). "Measuring self-regulated learning," in *Handbook of Self-Regulation*, eds P. Pintrich, M. Boekaerts, and M. Zeidner (Orlando, FL: Academic Press), 531–566. doi: 10.1016/b978-012109890-2/50045-7

Worth, K. (2019). *STEM in Early Childhood Education: How Science, Technology, Engineering, and Mathematics Strengthen Learning*. London: Routledge.

Zimmerman, B. J. (2000). "Attaining self-regulation: a social cognitive perspective," in *Handbook of Self-Regulation*, eds P. Pintrich, M. Boekaerts, and M. Zeidner (Orlando, FL: Academic Press).

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: historical background, methodological developments, and future prospects. *Am. Educ. Res. J.* 45, 166–183. doi: 10.3102/0002831207312909

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and mid middle school. *Dev. Rev.* 27, 172–223. doi: 10.1016/j.dr.2006.12.001

Zimmerman, C., and Croker, S. (2013). "Learning science through inquiry," in *Handbook of the Psychology of Science*, eds G. Feist and M. Gorman (New York, NY: Springer), 49–70.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Fridman, Eden and Spektor-Levy. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 July 2020
doi: 10.3389/fpsyg.2020.01665

Hands-On Exploration of Cubes' Floating and Sinking Benefits Children's Subsequent Buoyancy Predictions

Johanna E. van Schaik[1,2,3], Tessa Slim[3], Rooske K. Franse[4,5] and Maartje E. J. Raijmakers[1,2,4]*

[1] Educational Studies, Educational and Family Studies, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [2] LEARN! Research Institute, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [3] Educational Sciences, Education and Child Studies, Leiden University, Leiden, Netherlands, [4] Developmental Psychology, Department of Psychology, University of Amsterdam, Amsterdam, Netherlands, [5] NEMO Science Museum, Amsterdam, Netherlands

Children accrue experiences with buoyancy on a daily basis, yet research paints a mixed picture of children's buoyancy knowledge. Whereas children's predictions and explanations of the floating and the sinking of common objects are often based on a single feature (e.g., mass or facts), children's predictions of novel cubes reveal solution strategies based on mass and volume integrations. Correspondingly, category learning theory suggests that categories (e.g., floaters vs. sinkers) are easier to identify when items mainly vary from one another in the relevant defining features. For example, a set of cubes only varies in mass and volume and hence density, thereby being able to highlight the deterministic role of density when placed in water. Here we asked how item variation during hands-on exploration affects children's subsequent predictions and explanations of buoyancy. Kindergarteners and first-, second-, and third-grade children individually explored either a set of 10 systematically varied cubes (i.e., systematic condition; $n = 95$) or a set of 10 common objects (i.e., non-systematic condition; $n = 96$) in a water basin. Next, the children predicted the buoyancy of five new cubes and five new common objects one at a time. Subsequently, the children explained their predictions one subset at a time. The children in the systematic condition were more accurate in their predictions of the test cubes than the children in the non-systematic condition. Latent class regression analyses identified three cube prediction solution strategies. The children in the systematic condition were more likely to use a strategy in which buoyancy decisions were made based on an accurate integration of mass and volume, while the children in the non-systematic condition were more likely to use a strategy in which mass was given more predictive load than volume. A third strategy was characterized by guessing. Latent class analyses of the children's explanations revealed different explanation strategies, each appealing to several features, but as hypothesized, no clear condition differences were found. The findings indicate that even 5 min of exploration with systematically varied cubes can already help children use an advanced buoyancy prediction strategy. This provides evidence in favor of using category learning theory to inform early science education design.

Keywords: variation, buoyancy, science concepts, inquiry-based learning, category learning

OPEN ACCESS

Edited by:
Christopher Osterhaus, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Peter Adriaan Edelsbrunner, ETH Zürich, Switzerland
Dimitrios Stamovlasis, Aristotle University of Thessaloniki, Greece

****Correspondence:***
Johanna E. van Schaik
j.e.van.schaik@vu.nl

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *31 January 2020*
Accepted: *18 June 2020*
Published: *21 July 2020*

Citation:
van Schaik JE, Slim T, Franse RK and Raijmakers MEJ (2020) Hands-On Exploration of Cubes' Floating and Sinking Benefits Children's Subsequent Buoyancy Predictions. Front. Psychol. 11:1665. doi: 10.3389/fpsyg.2020.01665

INTRODUCTION

In many Dutch kindergarten classrooms, 4- to 6-year-old can explore what happens when they place objects in water. Buoyancy lesson plans for such water corners typically suggest offering children a collection of common objects to experiment with (Franse, 2007; Kemmers et al., 2007; Mundkur, 2015; Science NetLinks, 2020). While the use of common objects arguably helps children to relate their classroom investigations to their daily experiences, these objects vary over many features (e.g., texture, rigidity) that are unrelated to their buoyancy. This surface feature variation therefore might come at a cost: it could make the relevant feature (i.e., density[1]) harder to isolate. Here we investigated how variation between items during hands-on exploration influences children's buoyancy knowledge.

A solid object's buoyancy is determined by its density, which is its mass divided by its volume, relative to the density of the surrounding fluid. Water's density is approximately 1, and objects with a density of less than 1 float in water, while objects with a density greater than 1 sink. Density is a fundamental concept throughout the physical sciences (e.g., gases and fluids, weather). Furthermore, density is an example of proportional reasoning, entailing a multiplicative relation between quantities, which, in turn, is central to mathematical and scientific topics (Siegler, 1981; Boyer et al., 2008). Consequently, children's understanding of buoyancy and density has been the subject of developmental and educational research from multiple perspectives (Piaget and Inhelder, 1974; Smith et al., 1985; Kohn, 1993; Penner and Klahr, 1996; Schneider and Hardy, 2013).

From a very young age, children gain experience with buoyancy in different contexts and for a range of objects. This suggests that they are starting to construct knowledge of buoyancy before necessarily being able to verbalize this knowledge. Following Pine and Messer (1999) we will refer to these different types of knowledge as implicit and explicit knowledge, respectively. In this sense, a child's explicit knowledge necessitates a conceptual representation (although not necessarily a scientifically accurate one), while her implicit knowledge is not (yet) verbally accessible but can be used to (successfully) perform a task (Karmiloff-Smith, 1992; Pine and Messer, 1999, 2003; Messer et al., 2008). Buoyancy studies indeed indicate that both the nature of children's knowledge (i.e., implicit vs. explicit) at a given age as well as the types of items used to elicit children's buoyancy knowledge (i.e., common objects vs. novel cubes) paint conflicting pictures of children's developing understanding (Rappolt-Schlichtmann et al., 2007; Kloos, 2008; Franse et al., under review).

Children's predictions of whether common objects will float or sink (which could be made based on either implicit or explicit knowledge) are typically quite accurate. For example, in one study, kindergartners' and second-graders' prediction accuracies ranged between 60 and 90% (Rappolt-Schlichtmann et al., 2007). In a study by Franse et al., (under review) 4- to 12-year-old predictions of a boat, coin, ball, and pebble likewise averaged between 88 and 92% accuracy, with predictions significantly improving with age. Children's predictions were likely accurate because they were making use of known facts or experiences. Correspondingly, the majority of these children's explanations of why these objects float or sink (which requires explicit knowledge) relied on relevant facts, with mass explanations in second place (Franse et al., under review; see also Smith et al., 1985; Tenenbaum et al., 2004). Thus, young children's implicit and explicit knowledge in the context of common objects both seem often to be based on one feature (i.e., facts or mass).

Whereas the floating and the sinking of common objects can be learned through daily experiences, this is not the case for novel items. Surprisingly, when young children are presented unfamiliar cubes one at a time, which vary only in mass and volume and, hence, in density, they are relatively accurate in predicting which will float and which will sink (Kohn, 1993; Kloos, 2008). This effect was analyzed further in the aforementioned study by Franse et al. (under review). In this study, children compared items (cubes and common objects) one at a time to two reference cubes (a sinker and a floater) and predicted the item's buoyancy before subsequently explaining these predictions per set of items. The authors analyzed the way the cubes' features (e.g., mass and volume) were used by children to decide their buoyancy. The majority of the children's predictions were shown to utilize a solution strategy that was based on an integration of mass and volume. Children's use of these solution strategies, and the accuracy of the solution strategies, increased with age; while most 4- to 5-year-old guessed, the majority of children 6 years and older used integrative solution strategies, and the higher the age, the more advanced the solution strategy. Conversely, the same children's explanations for the buoyancy of the cubes were largely one-dimensional, with the majority of children relying only on mass (Franse et al., under review).

Taken together, across common objects and cubes, prediction accuracy increases with age. However, the predictions of objects are typically based on facts, while (the implicit) predictions of cubes reveal the children's emerging ability to use the relevant features of mass and volume by the end of kindergarten. Explanations regarding common objects and cubes are both largely one-dimensional and hence inaccurate. Improving (young) children's concepts of buoyancy would thus require both drawing attention toward the less apparent features of items (e.g., density) and a shift toward multidimensional thinking (Kuhn et al., 2008; Schneider and Hardy, 2013). Considering the integrative solution strategies found when children were simply presented with cubes without experimentation (Franse et al., under review) it would seem that allowing children to experiment with a set of cubes could further trigger children's attention to the integration of mass and volume (Franse et al., under review).

This proposition finds backing in category learning studies and theory. This research shows not only that humans are quick to identify relations across items by detecting the invariant features but also that such learning is dependent on the category structure (Erickson and Kruschke, 1998; Ashby and Ell, 2001; Sloutsky, 2003, 2010; Kloos, 2008; Kloos and Sloutsky, 2008;

[1]Note that Archimedes' law of buoyancy provides a more general and an adequate way of thinking about floating and sinking, that is, "the relevant feature is the volume of the water to replace relative to the weight of the object" instead of "density."

Goldwater et al., 2018). Categories that are dense, that is, that have multiple category-relevant features (e.g., shape and behavior of dogs) and few category-irrelevant features (e.g., floppy or upright ears), are easy to learn because of the abundance of defining, invariant features. Sparse categories, however, have multiple irrelevant features and only few relevant, invariant features, making them hard to learn without guidance (Kloos and Sloutsky, 2008). Scientific concepts, such as density, are often based on sparse categories due to their multi-contextuality. Hence, conceptual science learning requires being able to distill the harder-to-detect features across multiple instances. To do so, attention needs to be shifted from irrelevant features (e.g., texture) to relevant features (e.g., density). This selection is challenging and likely requires executive functioning systems (Sloutsky, 2010) which are notoriously prolonged in their development (Diamond, 2013). To facilitate this process, simplifying the learning input (i.e., making the instantiation of the category denser) should make it easier for children to focus on the relevant features simply because there are fewer irrelevant features to focus on. This stands in contrast to what is often done (or advocated) in educational practice. Buoyancy lesson plans often suggest using (a variety of) common objects (Franse, 2007; Kemmers et al., 2007; Mundkur, 2015; Science NetLinks, 2020). The motivation to select a range of objects could derive from a constructivist idea that experiences which build upon prior experiences facilitate knowledge construction (Campbell, 2015).

In summary, findings on children's knowledge of buoyancy are scattered and depend on how the knowledge was elicited. While children's predictions about the buoyancy of common objects are typically quite accurate, these are likely based on single features such as mass or facts instead of scientifically correct concepts. Intriguingly, children's predictions of novel cubes reveal solution strategies based on mass and volume integrations. Correspondingly, category learning theory suggests that when relevant features (e.g., density) are more apparent across items (e.g., cubes), these features are easier to learn. Following from this, we asked how item variation during hands-on experience with buoyancy affects children's implicit and explicit knowledge.

To this end, kindergarten and first-, second-, and third-grade children explored either a set of 10 systematically varied cubes (i.e., systematic condition) or a set of 10 common objects (i.e., non-systematic condition) in a water basin. Following the exploration, all children were presented with five new cubes and five new common objects one at a time in a randomized order. They were asked to predict the buoyancy of these new items and were subsequently asked for buoyancy explanations for the subsets of these items.

We hypothesized that children in the systematic condition would predict the buoyancy of the new cubes more accurately than children in the non-systematic condition. Particularly, based on the past cube prediction findings (Franse et al., under review), we expected to find that the systematic condition children would be more likely to acquire a solution strategy in which they integrate the mass and the volume of the cubes to make buoyancy decisions and that they do so to a higher degree of accuracy than the children in the non-systematic condition. We did not expect to see notable differences between the two conditions in their predictions of the new common objects because children are generally quite accurate at predicting these (Rappolt-Schlichtmann et al., 2007) and all children could still rely on their past experiences and known facts (Franse et al., under review).

With respect to children's explanations, this study was more exploratory in nature. Studies investigating children's understanding of buoyancy and other topics show a discrepancy between children's predictions, which are typically more advanced, and their explanations, which lag behind (Smith et al., 1985; Butts et al., 1993; Rappolt-Schlichtmann et al., 2007). Given this study's 5-min hands-on exploration session, it is unlikely that the systematic condition children would be able to directly explicate their experimental findings (Pine and Messer, 1999) let alone transfer this knowledge to common objects. Nevertheless, to assist this process, children in both conditions were offered the same regularly timed prompts to stimulate them to explicitly think about their experiments as well as status overviews to guide their experimentation progress (Lazonder and Harmsen, 2016). As such, children with higher prior knowledge to begin with (i.e., children in higher grades) might show some evidence of explaining the new items' buoyancies based on their exploration.

MATERIALS AND METHODS

Participants

Kindergartners (grades 1 and 2 in the Netherlands) were recruited at seven primary schools. Of the 151 who participated, the data of 40 (26%) were excluded due to the following reasons: confusing the terms floating and sinking (n = 2), corrupt video files (n = 4), experimenter errors (n = 19), stopping the experiment early (n = 3), teacher-reported non-normal development (n = 1), not sorting items during the exploration session (n = 2), and incorrectly sorting more than two of the items during the exploration session (n = 9)[2]. Hence, the data of 111 kindergartners (53 girls; **Table 1**) were included.

Ninety-three first-, second-, and third-graders (grades 3, 4, and 5 in the Netherlands) were recruited at a science museum. Of these, the data of 13 (14%) were excluded due to corrupt video files (n = 12) and parent-reported non-normal development (n = 1). This resulted in the inclusion of 80 children tested at the science museum; 28 first graders (14 girls), 23 second graders (14 girls), and 29 third graders (11 girls; **Table 1**).

The final sample consisted of 191 participants (92 girls). A total of 95 participated in the systematic condition and 96 participated in the non-systematic condition. Informed consent was acquired from the parents before participation. This study was approved by the local social science faculty's ethics committee (ECPW-2018/193 and ECPW-2018/204).

Design

Materials

Figure 1 depicts the items used in this study (see also **Supplementary Table S1**). The items were selected based on

[2]This was used as an exclusion criterion as a proxy of children's engagement during the exploration session and to ensure that the included children understood the difference between floating and sinking.

TABLE 1 | Number and age of the participants per grade and per condition.

Grade	Condition	Mean age	Minimum age	Maximum age	*n* (*n* girls)
Kindergarten	Systematic	60.15	49	76	53 (24)
	Non-systematic	66.79	50	82	58 (29)
1st grade	Systematic	81.14	75	93	14 (9)
	Non-systematic	82.50	74	96	14 (5)
2nd grade	Systematic	94.91	83	100	11 (8)
	Non-systematic	95.58	87	103	12 (6)
3rd grade	Systematic	107.94	100	121	17 (9)
	Non-systematic	105.92	101	116	12 (2)

FIGURE 1 | The item sets used in the study. **Top left:** exploration set of 10 cubes used in the systematic condition. **Bottom left:** exploration set of 10 common objects used in the non-systematic condition [boat (F), sink sieve (S), lava rock (S), ping pong ball (F), metal spoon (S), 2-Euro coin (S), marble (S), bouncy ball (F), wall plug (F), and coaster (F)]. **Right:** the 10 test set items used in both conditions (**top:** branch, candy tin, bobby pin, ceramic utensil rest, and floor ball). Note that each cube had a unique symbol (visible in the **top-left** picture) that the experimenter used to identify it, but there was no discernable pattern so that children could not use these to decipher a cube's buoyancy.

past findings and pilot study results. The cubes were made by 3D-printing hollow boxes and lids. These were subsequently filled with Styrofoam, lead pellets, and glue to achieve the desired masses, and the lids were super-glued shut.

Pilot Study

For the pilot study, a broad range of items were selected, drawing inspiration from previous studies and buoyancy lesson plans (Franse, 2007; Kemmers et al., 2007; Mundkur, 2015; Science NetLinks, 2020). Forty-three children between the ages of 4 and 10 years were tested individually at a science museum after receiving a signed informed consent from their accompanying guardian. Following an introduction with a water basin to ensure familiarity with the terms "floating" and "sinking," the experimenter handed items one at a time to the child and asked her to predict whether it would float or sink. Based on these data, a final selection was made, and for the objects, the materials, the shapes, and the presence of holes and hollow spaces were taken into account to create a diverse range. Several items that did not achieve the desired accuracy rates were replaced based on the accuracies of the other items and past studies' findings; hence, not all items in the final design were piloted. Where available, pilot or previous study prediction accuracy is included in **Supplementary Table S1**.

Introduction Set

To introduce the water basin and the terms "floating" and "sinking," the experimenter used a cork and a stone, respectively. As a visual reminder and to structure children's explorations, two laminated sheets were used, one depicting a cork floating in a water basin and the word "float," and the other depicting a stone at the bottom of a water basin and the word "sink" (adapted from Franse, 2007).

Exploration Sets

There were two exploration sets (**Figure 1**): one consisted of 10 common objects (used in non-systematic condition) and the other of 10 cubes (used in systematic condition). Both sets consisted of five floaters and five sinkers. Both sets had four items that were selected to be surprising for children based on the pilot study: two surprising floaters and two surprising sinkers. All of the systematic condition's exploration set cubes were turquoise.

Test Set

The test set was the same for both conditions and was used to measure children's predictions and explanations (**Figure 1**). It consisted of five common objects and five cubes. Of each set of five items (i.e., cubes or objects), two were selected to be easier and three were selected to be more difficult based on the pilot study. This resulted in four subsets: easy cubes (n = 2), easy common objects (the candy tin and the ceramic utensils rest; n = 2), difficult cubes (n = 3), and difficult common objects (the branch, bobby pin, and floor ball; n = 3). Of the easy items, one was a floater and one was a sinker. Of the difficult items, two were floaters and one was a sinker. All of the test set cubes were white.

Procedure

Introduction

Children were tested individually. The experimenter first introduced the water basin and asked what would happen to the cork (or stone, order counterbalanced) and placed the item in the water basin once the child had responded. The experimenter asked the child what had happened and then provided the corresponding definition [i.e., "If something stays on the water, we (indeed) call that floating" and "If something goes to the bottom of the water basin, we (indeed) call that sinking"]. The experimenter pointed to the corresponding laminated sheet and placed the object on it. After this was performed for both example items, these were removed and the experimenter brought out the exploration set of that child's condition.

Exploration

During the exploration session, the child was free to test the exploration set items in the water basin. The laminated float and sink sheets remained on the table for the child to use. Before starting, the child was encouraged to try to think about why some float and others sink. To further stimulate children to test all items and think about buoyancy, children were prompted by the experimenter at regular intervals. If after 30 s the child had not yet tried anything, the experimenter asked which item she wanted to put in the water first. After 3 min or if the child had indicated that she was finished before having tried all of the items, the experimenter asked if she had tried all of the items already. After 4 min or if the child indicated that she was finished, the experimenter asked if the child already knew why some float and some sink. After 5 min, the experimenter indicated that the time was up and removed the water basin and items.

Prediction

The experimenter placed the laminated float and sink sheets upright into two bins, creating a floating and a sinking bin. The experimenter then placed the test set items one by one in front of the child and asked whether the item would float or sink. If the child did not pick up the item, the experimenter instructed her to do so. Once the child made a prediction, she was instructed to place the item in the corresponding bin. This was repeated for all 10 items in a randomized order.

Explanation

The experimenter pulled the items belonging to one subset (i.e., the easy cubes, the easy common objects, the difficult cubes, or the difficult common objects) out of the bins, regardless of whether the child had sorted them correctly or not, and placed them on the table. The experimenter pointed to the items from the floating bin (if there were any) and the items from the sinking bin (if there were any) while saying, "You thought this/these would float and this/these would sink. Why do these float and do these sink?" After the child was done explaining, the experimenter asked, "Can you explain it some more?" This was repeated for each of the four subsets in a randomized order.

Measures

Predictions

Children's categorization of the test set items into the floating and the sinking bins was tallied in two ways: an accuracy score of incorrect (0) or correct (1) and a float–sink score of float (0) or sink (1).

Explanations

Children's explanations were transcribed per item subset (i.e., easy cubes, easy objects, difficult cubes, and difficult objects), resulting in four sets of transcriptions per participant. The transcriptions were coded using an amended version of the coding scheme used by Franse et al. (under review). The final codes belonged to six nominal categories (1: other, 2: fact, 3: mass, 4: volume, 5: material, 6: mass and volume or scientific; see **Supplementary Table S2**). All applicable codes were allocated per subset. For example, a child's explanation of the difficult objects could receive a 2, 4, and 5, for referring to a fact or experience, describing the item's volume, and naming a material, respectively. Twenty percent (n = 40) of the transcriptions were re-coded by a second coder. The two coders had an overlap of 90.54%. Disparities were resolved on the basis of the coding scheme.

The data were subsequently reformatted for the analyses. A child received a binary absence (0) or presence (1) score for each explanation category, depending on whether their explanation had not or had included that category of explanation, respectively. This was done separately for each subset. In other words, per explanation category and per item subset, each child had a score of 0 if their explanation did not include this category and a score of 1 if their explanation did.

Data Analysis

Predictions

To test whether the conditions affected the children's predictions of the test set items' buoyancy, a generalized linear mixed model (GLMM) for binomial data was performed with children's prediction accuracies of the 10 items as the dependent variables and item type (cubes and objects) and condition (systematic vs. non-systematic) and their interaction as the independent variables. The random intercept effects of items and participants were included in the model. The GLMM was performed in the R package lme4 (R Core Team, 2017). However, to test how the conditions influenced the degree to which children's predictions incorporated item features, we needed to capture the heterogeneity across children. For this, we used latent class analysis (LCA) techniques (Hickendorff et al., 2018) to identify patterns in children's buoyancy predictions across the different items (i.e., across the five cubes and five common objects). The LCAs were performed separately for the cubes and the objects because of their different features and are explained below.

Cubes

To test how children's predictions incorporated the relevant features (e.g., mass and volume) of the cubes, we used latent class regression analysis (LCRA) (Huang and Bandeen-Roche, 2004) following Franse et al. (under review). LCRAs have previously also been applied to cognitive development of other science concepts (Bouwmeester et al., 2004; Hofman et al., 2015). In the LCRAs, children's buoyancy predictions of the five cubes are modeled as a function of the mass and the volume of each cube (plus an intercept). Different combinations of the cubes' mass and volume in the buoyancy predictions can be captured by different classes, with each class being characterized by a regression equation. For example, one class might be characterized by a regression model that accurately integrates the mass and the volume of the cubes to predict buoyancy, and another class' buoyancy predictions might best be modeled by a higher predictive load of mass than volume. Thus, each class identifies a solution strategy, namely, the way mass and volume were combined to decide items' buoyancies, and hence the terms class and solution strategy can be used synonymously when interpreting the results. The logarithms of the cubes' mass and volume were used as the predictors in the regression models because people's perception of mass and volume operates on a logarithmic scale (Jones, 1986). Since children's buoyancy predictions (i.e., the dependent variables) were scored as float (0) or sink (1), the models indicate the probability of a child predicting that a cube will sink.

One, two, and three class models were fit to the children's float–sink predictions, including regression terms b_{mass}, b_{volume}, and the intercept, a. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to select the optimal number of classes. Subsequent constrained models (with fewer regression terms) were compared using likelihood ratio tests based on Pearson's χ^2. Once the most parsimonious model was selected, the posterior probabilities were computed, indicating, for each child, the probability that the child belongs to each class based on their buoyancy predictions. The highest posterior probability was used to assign the child to a class.

Objects

Since the objects did not differ systematically in mass and volume, latent class analyses (McCutcheon, 1987) were used to detect underlying patterns in the dependent variables, namely, the accuracy of children's buoyancy predictions across the five cubes. LCAs were fit with different numbers of classes, and models were compared using AIC, BIC, and bootstrapped model fit likelihood ratios (Hickendorff et al., 2018). The selected n-class model was subsequently made more parsimonious by constraining responses within classes and then between classes to be equal, freeing up degrees of freedom. LC(R)As were conducted using DepMixS4 in R (Visser and Speekenbrink, 2010) and MPlus (Muthén and Muthén, 2017).

Condition and Grade Effects

Following the "three-step approach" (Hickendorff et al., 2018) separate logistic regressions were carried out to test the effect of condition and grade on children's cube class and object class memberships.

Explanations

Children's explanations were also modeled using LCAs. In this manner, we could detect underlying patterns in the types of explanations children gave [i.e., the dependent variables were the absence (0) or presence (1) of each explanation category]. This is arguably more informative than examining the highest

achieved explanation or the amount of explanation types given because it has the potential to reveal (combinations in) the types of features children's explicit concepts of buoyancy rely on without defining these *a priori*. LCAs were again performed separately for the cubes and the objects, following the same procedure as for the predictions. The classes resulting from the LCAs differ in the degree to which class members were likely to use an explanation type per subset of items After the LCAs, separate logistic regressions were again performed on the cube and the object class memberships to test the effects of condition and grade.

Predictions vs. Explanations

To investigate whether there was a relation between children's implicit and explicit responses, logistic regressions were carried out separately for the two item types. Children's implicit class membership as well as grade were used to predict children's explicit class membership.

RESULTS

Buoyancy Predictions

Children's buoyancy prediction accuracy was compared across items between the two conditions using a GLMM (see **Figure 2** for summed prediction accuracies per item type and **Supplementary Figures S1, S2** for histograms). **Table 2** shows the fixed and random effects of the GLMM (see **Supplementary Table S3** for the correlations between the independent variables). The main effect of condition indicates that children in the non-systematic condition had a 0.35 lower odds of accurately predicting the buoyancy of an item than children in the systematic condition. The interaction between condition and item type, as displayed in **Figure 2**, indicates that children in the systematic condition were more accurate than children in the non-systematic condition, particularly on the cubes. This suggests that the systematic condition benefitted children's prediction of the test set cubes while not greatly affecting children's predictions of objects, whereas the non-systematic condition did not seem to benefit children's object predictions relative to the systematic condition.

Cubes

To give an optimal description of individual differences in children's buoyancy decisions about cubes, LCRA models with one to four classes were fit to children's predictions of sinking (**Table 3**). The best-fitting, most parsimonious model was a three-class model, with two classes that included mass and volume predictors ("mass and volume" class: $b_{mass} = -44.5$, $b_{volume} = 49.71$, $a = -22.19$, 29% of the children; "MASS and volume class": $b_{mass} = -1.84$, $b_{volume} = 3.68$, $a = -5.83$, 35% of the children) and one class that only included an intercept ("residual" class: $a = -0.31$, 36% of the children).

The selected LCRA model is depicted in **Figure 3**, in which for each of the three classes (i.e., solution strategies), the probability of predicting that a cube will sink is plotted as a function of mass and volume. This is done separately in five graphs for the five constant volumes of the tested cubes (i.e., 125, 166, 216, 27, and 275 cm^3), although note that the five graphs depict the same model. The black vertical line denotes the density of water (1 g/cm^3), such that hypothetical cubes of constant volume with a mass to the left of this line would float and those of constant volume with a mass to the right of this line would sink. Take, for example, the volume of 125 cm^3; the logistic curve of the "mass and volume" solution strategy closely aligns with the black density line, indicating that these children are modeled as correctly switching from predicting that a cube will float to predicting that it will sink when cubes have a density of around 1 g/cm^3. This suggests that they quite accurately make a buoyancy decision based on mass and volume. For the tested cube of 125 cm^3 and 37.5 g, this model therefore predicts that children will say that the cube will float (asterisk), which closely aligns with the observed buoyancy predictions (dot with error lines) of the children who were assigned to this class based on the posterior probabilities (see below). The "MASS and volume" logistic curve is less steep, indicating more uncertainty, and falls largely to the left of the density line. This implies that, at a lower mass, these children are already switching to predicting sinking, such that they seem to be over-compensating for mass. Indeed the model's prediction of the probability that children will say that the tested cube (which floats) will sink is 0.2 (asterisk), overlapping with the observed buoyancy predictions for this cube (dot with error lines). Lastly, the "residual" class only has an intercept and therefore no regression slope. Overall, their observed (dot with error lines) and modeled (asterisk) buoyancy predictions are around chance, suggesting a guessing strategy.

Children were assigned to one of the three classes based on the LCRA's posterior probabilities, which indicate for each child the probability of that child belonging to each class based on the child's actual buoyancy predictions. Next, a multinomial logistic regression revealed that this class membership could be predicted by condition and grade [Cox and Snell $R^2 = 0.37$, $\chi^2(6) = 89.17$, $p < 0.001$; **Figure 4**]. Relative to the "MASS and volume" class, children in the systematic compared to the non-systematic condition had an odds of 7.60 (95%CI: 3.16–18.27) to belong to the "mass and volume" class, $b = 2.03$, $p < 0.001$. In other words, children in the systematic condition were more likely to belong to the "mass and volume" class than the "MASS and volume class," while the opposite holds for children in the non-systematic condition. Children in the systematic condition compared to the non-systematic condition also had an odds of 9.97 (95%CI: 1.73–57.47) of belonging to the "residual" class relative to the "MASS and volume" class ($b = 2.30$, $p = 0.01$). As grade increases, the odds (0.06, 95% CI: 0.01–0.44) of belonging to the "residual" class relative to the "MASS and volume" class decreases ($b = -2.81$, $p = 0.005$). There was no significant interaction between condition and grade.

Objects

To give an optimal description of individual differences in children's buoyancy decisions about objects, LCAs were performed on the prediction accuracy of the five test objects. One, two, and three class models were compared, leading to

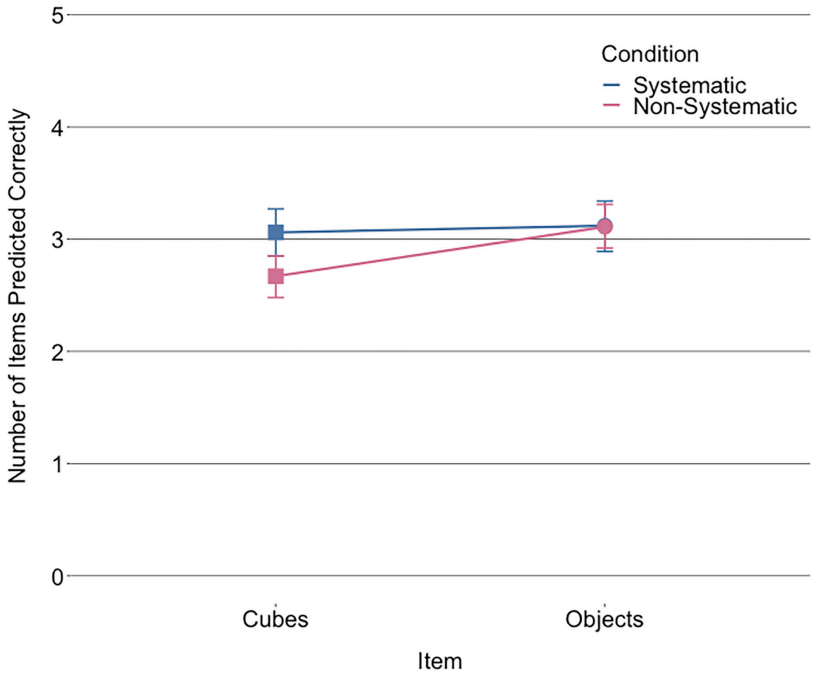


FIGURE 2 | Mean number of accurate predictions of the cubes and objects at test per condition. Error bars denote 95% confidence intervals.

TABLE 2 | Generalized linear mixed model statistics on the prediction accuracies.

	Fixed effects				Random effects (SD)	
	Beta	**SE**	***z*-value**	**Significance**	**Item**	**Participant**
Intercept	0.50	0.26	1.96	0.051	0.28 (0.53)	0.03 (0.17)
Condition	−0.35	0.14	−2.54	0.011		
Item type	0.03	0.36	0.08	0.93		
Condition * Item type	0.35	0.19	1.82	0.069		

SE, standard error; SD, standard deviation.

TABLE 3 | Fit statistics for latent class regression models of cube predictions.

Model	Classes	LR	Parameters	AIC	BIC	*p* (LR)
$\beta_{mass} + \beta_{vol}$	1	−611.49	3	1,228.98	1,243.56	
β_{mass}	1	−628.81	2	1,261.63	1,271.35	<0.001
β_{vol}	1	−653.52	2	1,311.03	1,320.74	<0.001
1c: $\beta_{mass} + \beta_{vol}$, 1c: α	2	−587.96	5	1,185.92	1,210.23	
1c: β_{vol},1c: α	2	−648.47	4	1,304.95	1,324.40	<0.001
1c: β_{mass}, 1c: α	2	−612.68	4	1,233.37	1,252.82	<0.001
*2c: $\beta_{mass} + \beta_{vol}$, 1c: α	3	−576.72	9	1,171.44	1,215.20	
1c: $\beta_{mass} + \beta_{vol}$, 1c: β_{mass}, 1c: α	3	−582.72	8	1,181.45	1,220.34	0.001
1c: $\beta_{mass} + \beta_{vol}$, 1c: β_{vol}, 1c: α	3	−582.52	8	1,181.03	1,219.93	0.001
3c: $\beta_{mass} + \beta_{vol}$, 1c: α	4	−569.88	13	1,165.76	1,228.96	

*Model, regression terms per class; 1c, one class' regression model included these terms; 2c, two classes' regression models included these terms; 3c, three classes' regression models included these terms; *, selected model; Classes, number of classes; LR, log likelihood ratio; AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; p (LR), p-value of likelihood ratio Pearson's χ^2.*

the selection of the two-class model (**Table 4**). This model was subsequently constrained. In the final model (LR = −583.80; *df* = 7; AIC = 1,181.60; BIC = 1,204.37), the log and the bobby pin were constrained to be equal within both classes, the floor ball and candy tin were constrained to be equal within one class, and the ceramic was constrained to be

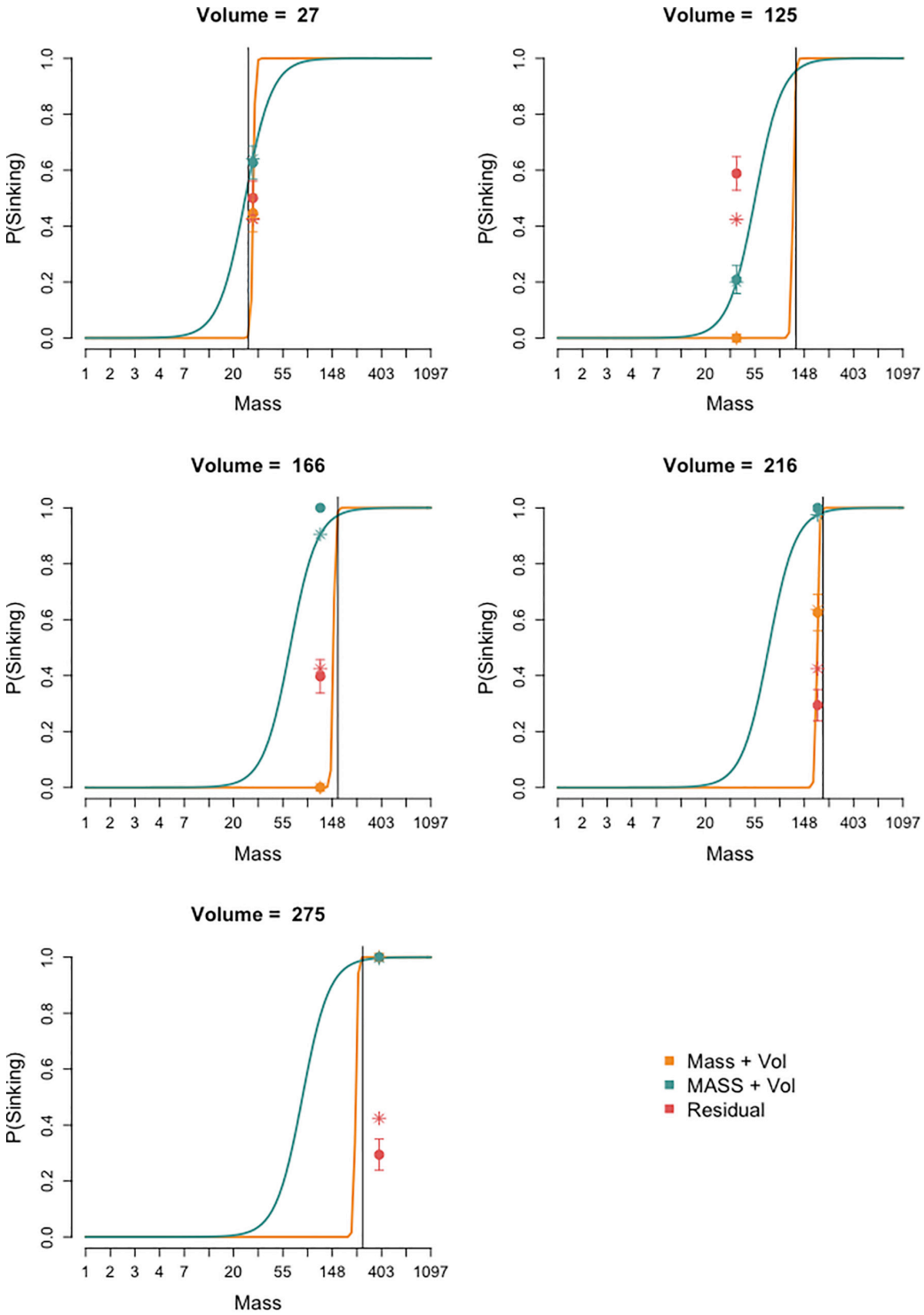


FIGURE 3 | The three classes (depicted in different colors) found in the optimal latent class regression analysis of children's predictions of cubes' sinking. The probability of predicting that a cube will sink is depicted in each of the five graphs as a function of mass; each graph depicts the models for a different constant volume (i.e., the volumes of the five tested cubes: cube 11: v = 125, m = 37.5; cube 12: v = 166, m = 116.5; cube 13: v = 216, m = 194.5; cube 14: v = 27, m = 30; and cube 15: v = 275, m = 384.5). The black vertical line denotes the density of water (1 g/cm^3); hence, cubes of constant volume with a mass to the left of this line would float and cubes of constant volume with a mass to the right of this line would sink. Each graph displays the modeled classes (curves as a function of mass and asterisks as a function of the predicted value for the mass of the tested cube) and the observed values (circles with standard errors). Classes "Mass and Volume" and "MASS and Volume" include all regression terms (hence, curves, and asterisks), while class "Residual" only included an intercept (hence, only asterisks). Note that the x-axis scale is logarithmic because logarithms of the cubes' mass and volume were used.

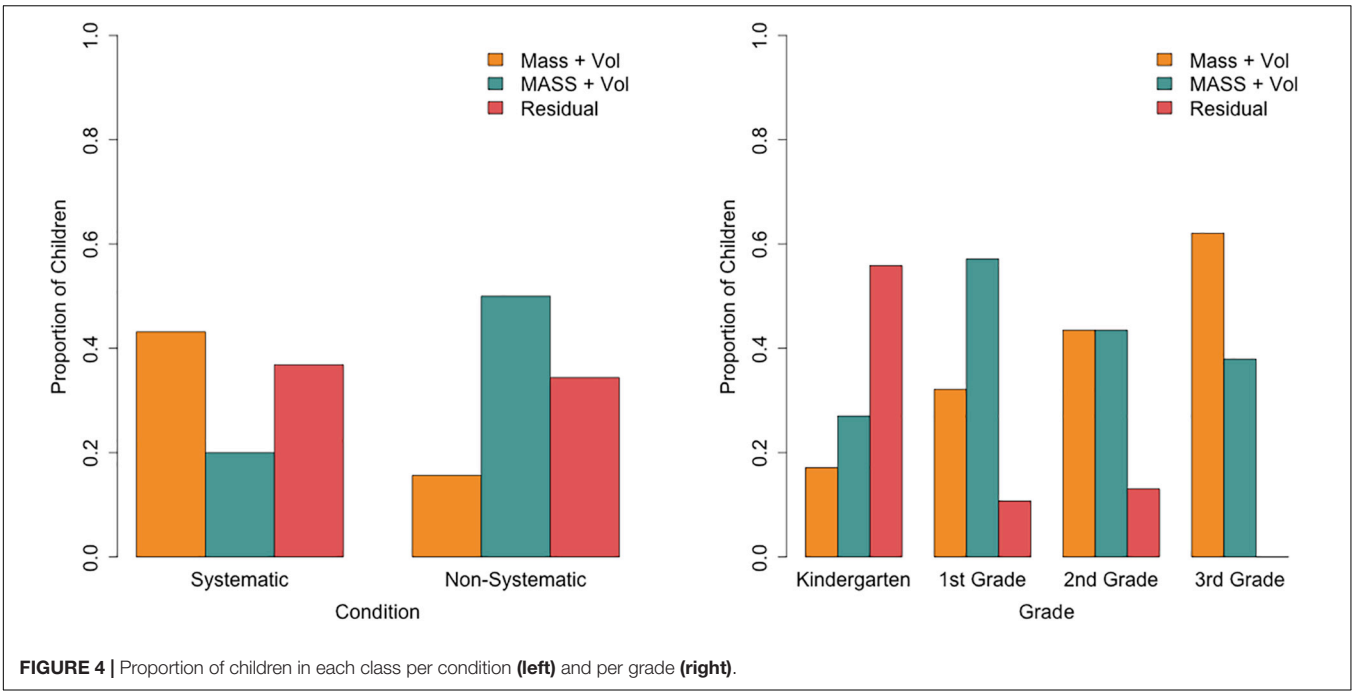


FIGURE 4 | Proportion of children in each class per condition **(left)** and per grade **(right)**.

TABLE 4 | Fit statistics for latent class models of common object predictions.

Model	LR	Df	AIC	BIC	aBIC	p^a (LR)	p^a (PLR)	Entropy
1 class	−609.91	5	1,229.81	1,246.07	1,230.24	0.00	na	
*2 class	−583.03	11	1,188.07	1,223.84	1,189.00	0.08	<0.001	0.56
3 class	−577.92	17	1,189.85	1,245.14	1,191.29	0.35	0.235	0.68

**, selected model; LR, log likelihood ratio; df, degrees of freedom; AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; aBIC, adjusted Bayesian Information Criterion; p (LR), p-value of model fit likelihood ratio Pearson's χ^2; p (PLR), p-value of parametric likelihood ratio for n-1 vs. n classes. [a]Bootstrapped values.*

equal across classes, indicating that it did not distinguish class membership.

Figure 5 displays the conditional probabilities of accurately predicting an object's buoyancy for both classes. One class (54% of the children) performs poorly on two objects in which the mass is misleading (i.e., the branch floats but is relatively heavy and the bobby pin sinks but is relatively light) but performs better when predicting that relatively light items will float (i.e., the floor ball and the candy tin). The second class (46% of the children) performs well on the misleading objects (i.e., branch and bobby pin) but performs around chance on the light items (i.e., the ball that has holes and the candy tin with air inside that is made of metal). These two classes are tentatively interpreted as a class that seems to use primarily object mass to predict buoyancy and a class that makes object-specific predictions, respectively.

Limited evidence (Cox and Snell R^2 = 0.07) was found for predictive effects of grade and condition on class membership as tested with a binary logistic regression (**Figure 6**). Children in the systematic condition had a higher odds (2.16, 95%CI: 1.18–3.94) than children in the non-systematic condition of belonging to the "object-specific" class relative to the "mass-based" class (b = 0.77, p = 0.012). As grade increases, the probability of belonging to the "object-specific" class also decreases (b = −0.57, p = 0.014). The interaction between condition and grade did not reach significance.

Explanations

Cubes

The coded explanation categories were fit with LCAs to detect patterns in which types and combinations of buoyancy explanations children gave. Models with increasing numbers of classes were compared. Models with more than four classes did not result in stable solutions, leading to the selection of the four-class model (**Supplementary Table S4**). This model was subsequently constrained to make it more parsimonious. Constraints were applied across easy and difficult cubes, starting with the simpler categories (e.g., other, then fact) and subsequently to the more complex categories (e.g., mass and volume or scientific). Constraints were applied within one class and, subsequently, within increasing numbers of classes. Lastly, constraints across classes were tested.

In the final model (LR = −693.22; df = 30; AIC = 1,446.44; BIC = 1,544.01), the other explanation category was constrained across easy and difficult cube explanations for three classes. Fact was constrained to be equal between easy and difficult cube explanations for all classes, as was material which was even constrained across classes, indicating that it did not distinguish

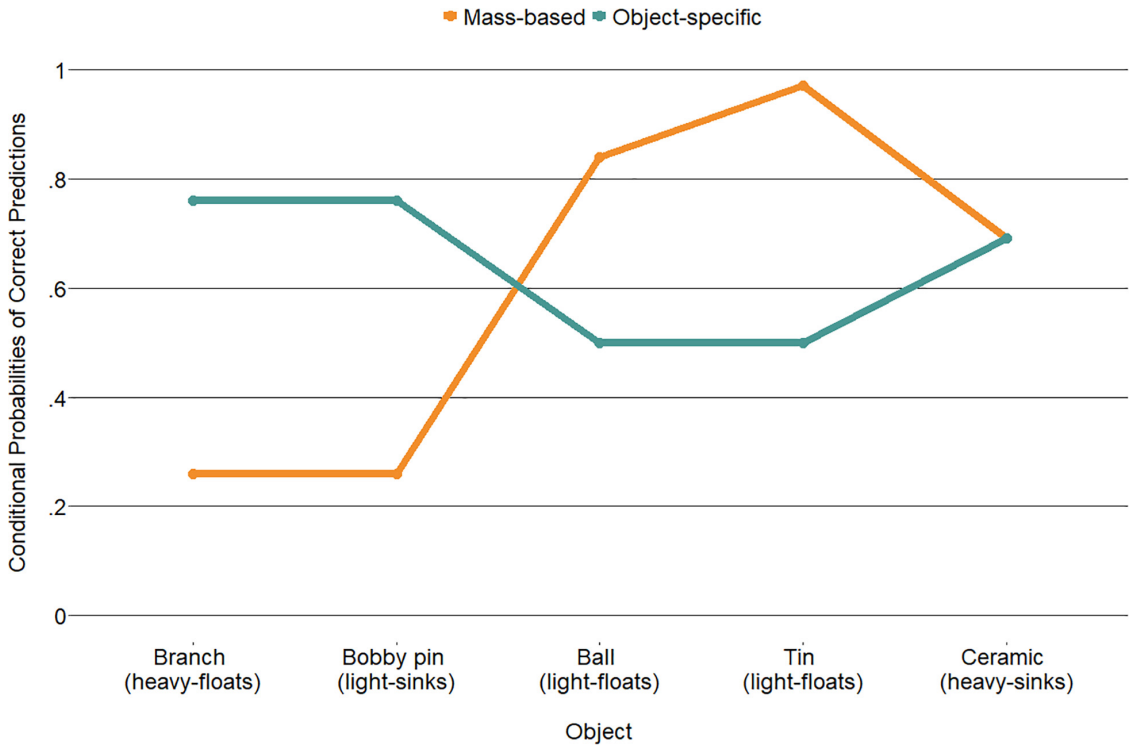


FIGURE 5 | Latent class analysis probabilities of prediction accuracy for the two classes across the five objects.

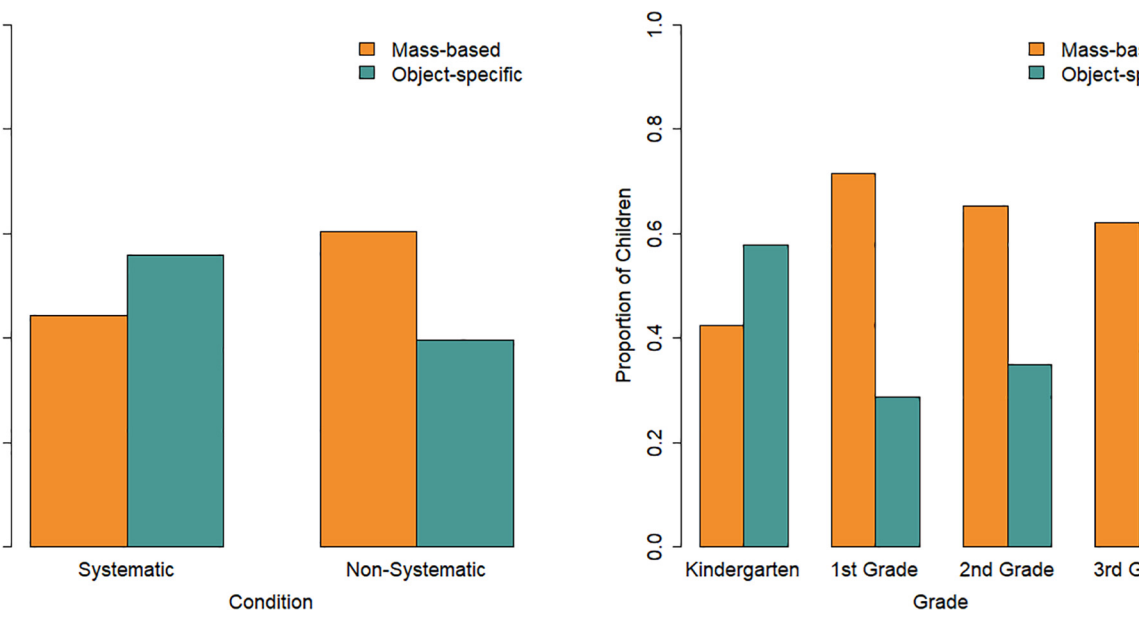


FIGURE 6 | Proportion of children in each object prediction class per condition **(left)** and per grade **(right)**.

classes. Mass was only constrained to be equal across easy and difficult explanations for one class, while this could be done for volume in three classes. Finally, the mass and volume or scientific category was constrained in two classes.

The four classes (**Figure 7**) were interpreted based on the explanation categories class members were most likely to use. The "M" class (47% of the children) was found to have a high probability of appealing to the cubes' mass for both the easy and

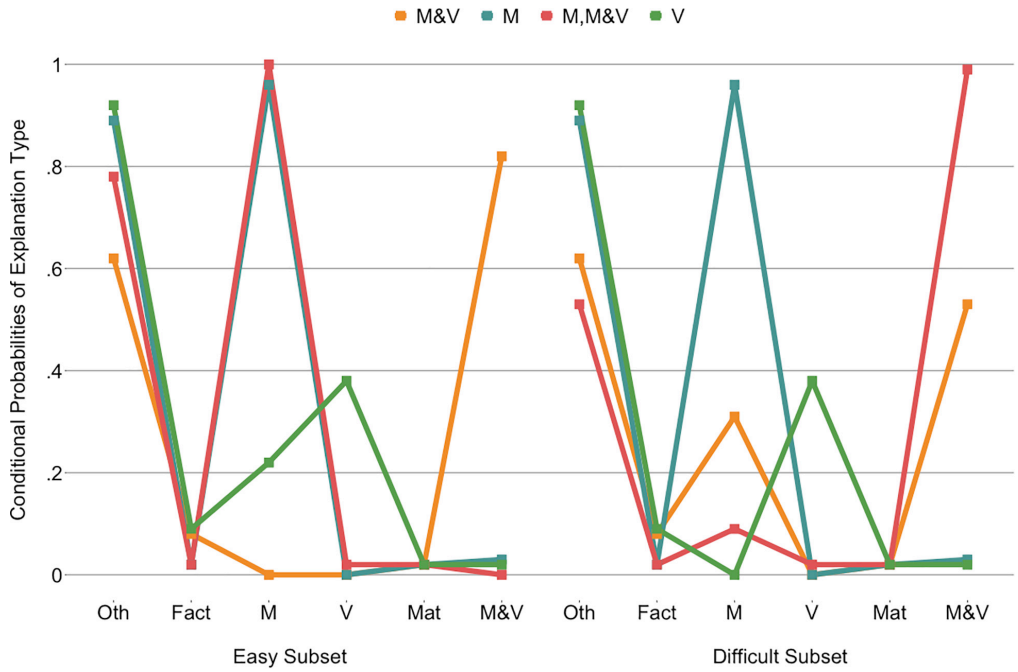


FIGURE 7 | Latent class analysis probabilities of explanation types for the four classes across the two cube subsets. Oth, other; M, mass; V, volume; Mat, material; M&V, mass and volume, scientific.

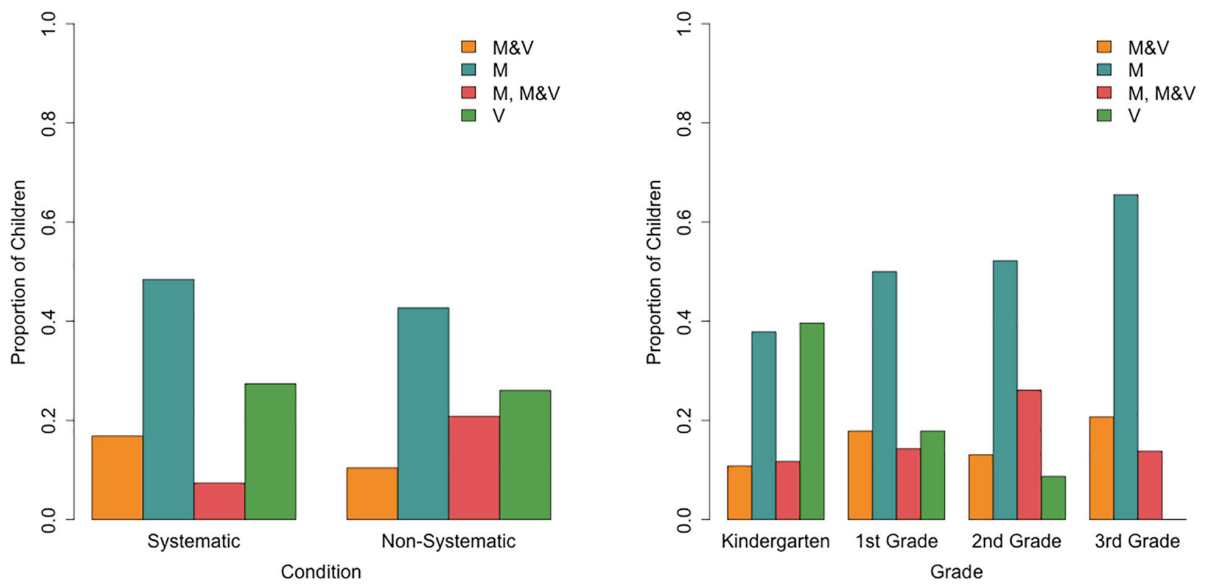


FIGURE 8 | Proportion of children in each cube explanation class per condition **(left)** and per grade **(right)**.

the difficult cube sets, as well as providing explanations in the other category. The "V" class (25% of the children) primarily had explanations that fall into the other category but appealed to volume more than the other classes did. The "M, M&V" class (14% of the children) used mainly mass in their explanations for the easy subset and mass-and-volume or scientific explanations for the difficult subset. The "M&V" class (14% of the children) had an inversed pattern, although their explanations on the difficult subset were more mixed, including both mass and mass-and-volume or scientific explanations. Notably, all classes gave explanations falling in the other category, although this was most prominent for the "M" and "V" classes.

A multinomial logistic regression with the "M" class as the reference class provided limited evidence for effects

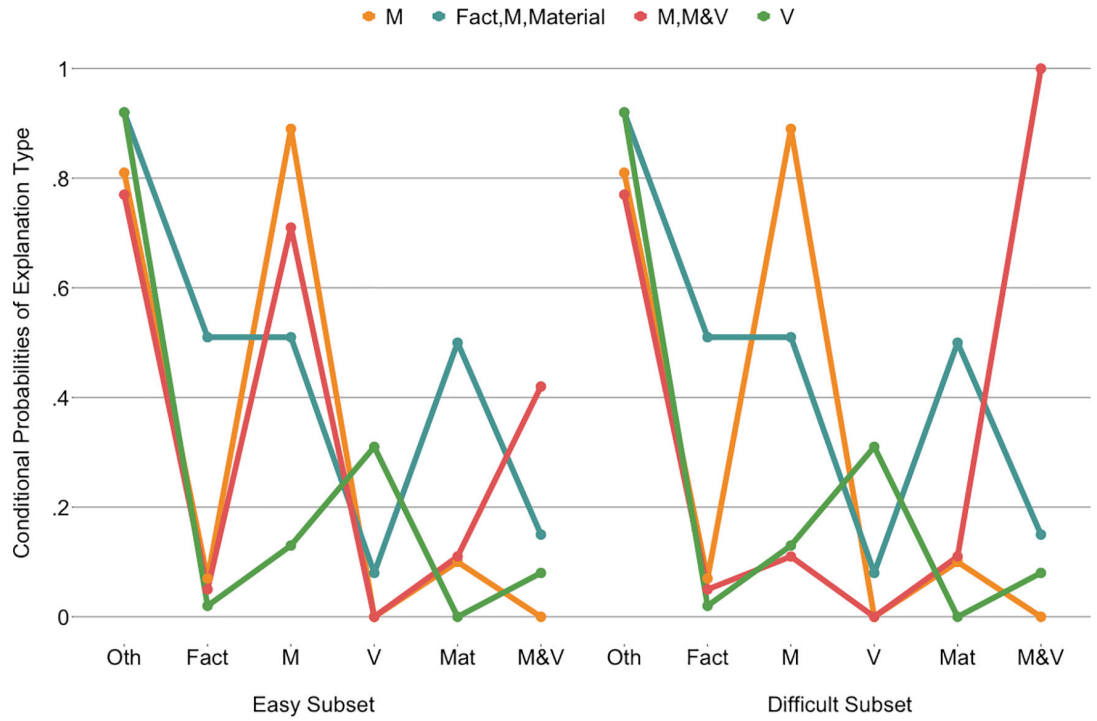


FIGURE 9 | Latent class analysis probabilities of explanation types for the four classes across the two object subsets. Oth, other; M, mass; V, volume; Mat, material; M&V, mass and volume, scientific.

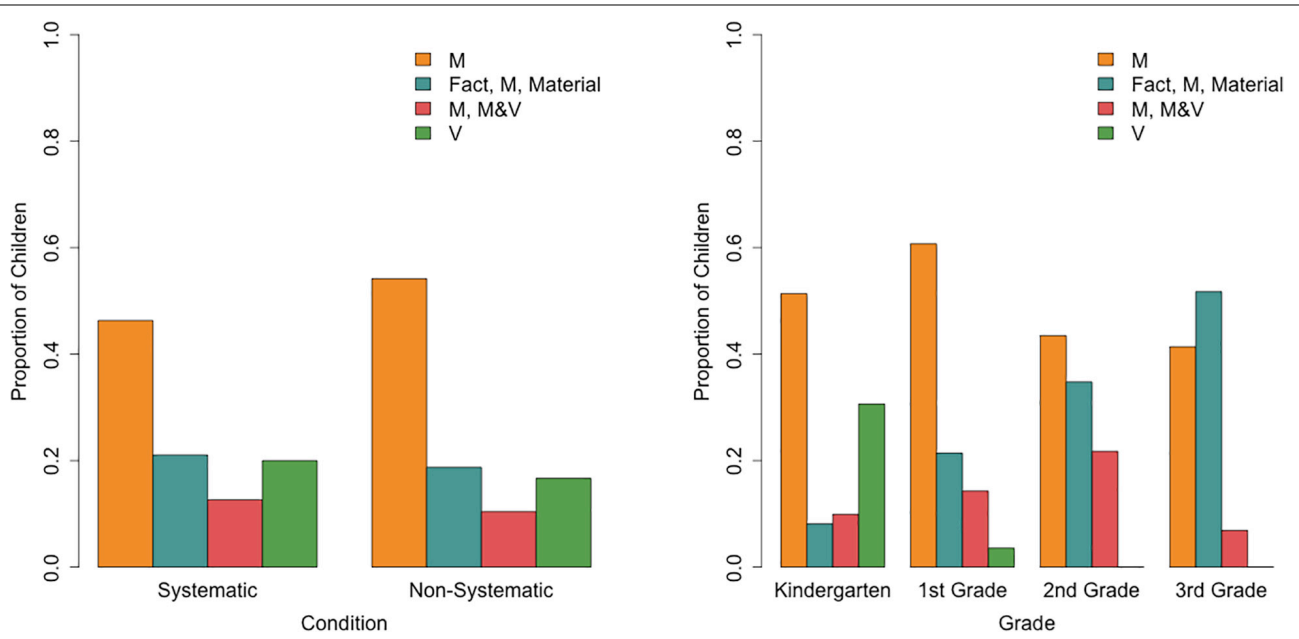


FIGURE 10 | Proportion of children in each object explanation class per condition **(left)** and per grade **(right)**.

of condition and grade on class membership [Cox and Snell R^2 = 0.19, $\chi^2(9)$ = 40.39, $p <$ 0.001; **Figure 8**]. Relative to the "M" class, the odds of belonging to the "M, M&V" class was lower for children in the systematic condition (0.32, 95%CI: 0.12–0.86) than in the non-systematic condition (b = -1.14, p = 0.023). As grade increased, the odds (0.37, 95%CI: 0.17–0.81) of belonging to the "V" class decreased relative to the "M" class (b = −1.00,

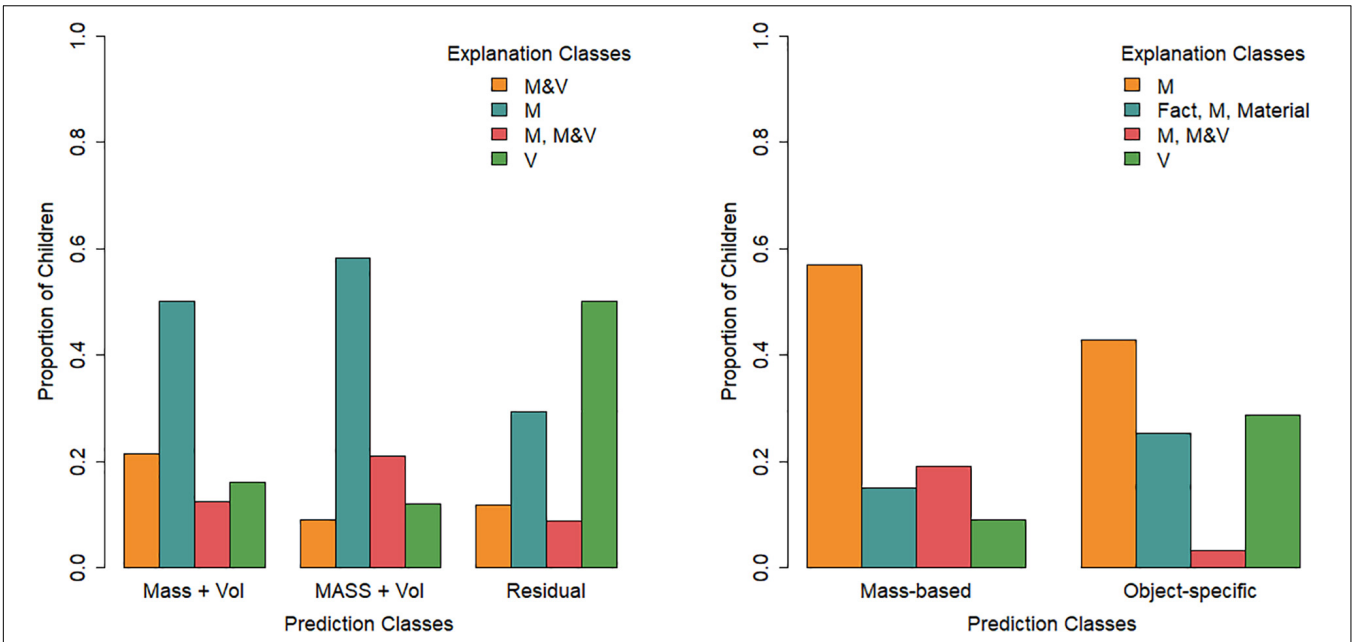


FIGURE 11 | Proportion of children in each explanation class per prediction class for the cubes **(left)** and the objects **(right)**.

p = 0.012). There was no significant interaction effect between grade and condition.

Objects

LCAs to detect patterns in children's explanations of the objects were run in the same way as those for the cubes (**Supplementary Table S5**). Note that although the prediction accuracies (see **Figure 5**) did not necessarily reflect the *a priori* classification of the objects into easy and difficult subsets, explanations were asked for an entire subset at a time and could therefore not be analyzed separately per object. Models with more than five classes did not identify. The four-class model was selected as the BIC was lower than for the five-class model and for ease of interpretation because of the qualitative overlaps with the selected cube explanation model.

This model was constrained in the same theory-driven manner as the cube model. In the final model (LR = −903.89; df = 29; AIC = 1,865.79; BIC = 1,960.10), the other, fact, material, and volume categories were constrained across easy and difficult subsets for all classes. This was also the case for the mass and the mass and volume, scientific categories for three of the classes. No categories were constrained across classes. The classes of the final model (**Figure 9**) resembled those of the cube model: an "M" class (44% of the children), an "M, M&V" class (13% of the children), and a "V" class (18% of the children). The most notable difference was the presence of a class that appealed to facts and materials in addition to mass across easy and difficult objects, the "facts, M, material" class (25% of the children).

The multinomial logistic regression with the "M" class as the reference class revealed only one significant effect of grade on class membership [Cox and Snell R^2 = 0.27, $\chi^2(9)$ = 59.86, $p <$ 0.001; **Figure 10**]. As grade increases, the odds (1.92, 95%CI: 1.19–3.08) of belonging to the "fact, M, material" class increases relative to the "M" class (b = 0.65, p = 0.007).

Comparing Predictions and Explanations

To test whether there was a relation between children's predictions of cubes and their subsequent explanations (**Figure 11**) beyond grade effects, we performed a stepwise multinomial logistic regression in which grade and, subsequently, prediction class membership were added in separate steps. Adding prediction class membership did significantly improve the model [$\chi^2(6)$ = 17.14, p = 0.009]. In this model [Cox and Snell = 0.23, $\chi^2(9)$ = 48.51, $p <$ 0.001], the children in the "MASS and volume" prediction class had a lower odds of belonging to the "V" explanation class relative to the "M" explanation class as compared to the "residual" prediction class. In other words, children who used the guessing solution strategy to predict the cubes' buoyancy were more likely than the "MASS and volume" prediction class to belong to the "V" explanation class.

The same analysis was done for the object classes. The addition of the prediction classes improved the model of children's explanation classes beyond the effects of grade [$\chi^2(3)$ = 24.38, p<0.001]. Here [Cox and Snell = 0.34, $\chi^2(6)$ = 80.28, $p <$ 0.001] the prediction classes significantly predicted explanation class membership, with the "M" class as the reference class. Children in the "mass-based" prediction class had lower odds (0.23, 95%CI: 0.12–0.70, b = −1.25 p = 0.006) of belonging to the "fact, M, material" explanation class, higher odds (4.29, 95%CI: 1.17–15.77, b = 1.46, p = 0.028) of belonging to the "M, M&V" explanation class, and lower odds (0.35, 95%CI: 0.14–0.86, b = −1.05 p = 0.022) of belonging to the "V" explanation class as compared to the "object-specific" prediction class.

DISCUSSION

The current study compared the effect of children's hands-on exploration of buoyancy using cubes (i.e., systematic condition) vs. common objects (i.e., non-systematic condition) on children's subsequent predictions and explanations of new cubes' and common objects' buoyancies. Children's responses were investigated using categorical latent variable models to account for individual differences.

Predicting and Explaining Cubes

The latent class regression analyses of children's cube buoyancy predictions identified three classes corresponding to three solution strategies. These three strategies replicate those identified by Franse et al. (under review). While in one class the children's predictions were best modeled as guessing, two two-dimensional solution strategies were identified. The "mass and volume" solution strategy was characterized by a fairly accurate decision threshold, as indicated by a switch between predicting floating vs. sinking near a density of 1. This threshold was also modeled as being quite certain, as visible in the steepness of the decision curve (see **Figure 3**). The "MASS and volume" solution strategy was characterized as being less accurate as the threshold for predicting sinking was at a lower mass than would actually be the case for a given volume. Additionally, this solution strategy was also characterized by more uncertainty, seen in the more gradual slope from predicting floating to predicting sinking. Thus, both solution strategies made float–sink decisions based on the cubes' relevant features (i.e., integrating mass and volume), although they differed in accuracy (i.e., how to relate the features to a decision; *cf*, Kuhn et al., 2008) and uncertainty (i.e., the strictness of the float–sink threshold).

Importantly, the proportion of children across the three classes differed between the two conditions. Children in the systematic condition were more likely to be in the "mass and volume" class than the "MASS and volume" class, while this was the other way around for those in the non-systematic condition. There was no evidence for an interaction between condition and grade, suggesting that the systematic condition was not more effective in one grade than another. These findings indicate that the 5 min of testing cubes in the water basin was effective in helping children to use a more accurate and certain solution strategy when faced with novel cubes.

Given that children in higher grades were also more likely to use the more advanced solution strategy, it is not the case that the systematic condition's experience was *necessary* for using the "mass and volume" solution strategy. The effectivity of the systematic condition instead seems to lie in encouraging the switch to a more advanced integrative solution strategy. Since the two conditions differed in the proportion of children in one or the other integrative solution strategy but had similar proportions of children using the guessing strategy, it could further be postulated that this switch only occurs once a child already uses an integrative solution strategy but a less advanced one (e.g., the "MASS and volume" strategy; see also Siegler and Chen, 2008). This would need to be tested further by examining children's strategy use before and after hands-on experience and by measuring solution strategy changes over a longer period of time.

Four classes of explanations were identified with latent class analyses (LCAs): "M&V," "M," "M, M&V," and "V." Grade effects revealed that children in higher grades were less likely to belong to the "V" class, which was characterized by mainly other and some mass and some volume category explanations, and more likely to belong to the "M" class, which only used other and mass categories. However, as expected given the limited duration and scope of the exploration phase, the effects of condition did not strongly carry over to children's explanations of cubes. Only one condition effect was found; the "M, M&V" class membership was lower for systematic condition children than non-systematic condition children relative to the "M" class, yet the implications of this difference are unclear since several classes used mass and mass and volume explanations.

While by third grade all children were using a two-dimensional prediction solution strategy for the cubes, the majority of these children were explaining the cubes' buoyancies primarily by referring to their masses. This suggests a discrepancy between predictions and explanations as found across past studies (Smith et al., 1985; Butts et al., 1993). Indeed the only significant predictive effect of prediction strategy on explanation strategy, beyond the effects of grade, indicated a relation between using the guessing prediction strategy and belonging to the "V" explanation class, which was characterized by primarily other category explanations. The absence of more clear relations suggests that it is difficult for children to verbalize their use of mass and volume or density (see also Dienes and Perner, 1999).

Predicting and Explaining Common Objects

The accuracies of children's predictions of the common objects were analyzed with LCAs to detect underlying patterns across predictions. Two classes were identified. In one, dubbed "mass-based," the lighter objects had a higher probability of being accurately predicted to float, while a heavy floater and a light sinker were less accurately predicted. The second, "object-specific" class, had higher probabilities of correctly predicting the heavy floater and light sinker but were around chance at predicting the lighter objects. The "mass-based" class membership was slightly higher in the non-systematic than in the systematic condition.

Although it is not possible to determine which features children used to make these predictions, the most parsimonious interpretation would be that children based their explanations on a simple, one-dimensional strategy. The "mass-based" class responses could indeed be derived using the one-dimensional rule that heavy items sink and light items float. The "object-specific" class is trickier to interpret. These children could have used facts or object-specific past experiences to be more likely to correctly predict that a branch floats and a bobby pin sinks. The guessing behavior on the other items could stem from multiple sources of information; the ball was lightweight but had holes, reflecting the common idea that items with holes sink, and the

tin was made of metal which typically sinks but was filled with air which typically causes objects to float (Hardy et al., 2006).

The relation between children's common object predictions and explanations might help to further understand the prediction strategies. Four object explanation classes were identified: "M," "fact, M, material," "M, M&V," and "V." Whereas the majority of the children in the "mass-based" prediction class went on to belong to the "M" explanation class, the children in the "object-specific" prediction class were more distributed over the explanation classes. Notably, compared to the "mass-based" children, more "object-specific" children later belonged to the "fact, M, material" class or the "V" class which were both highly likely to also mention things in the other category. It could thus be that the "object-specific" prediction class is not so homogenous in how they came to the predictions but instead used several sources of information. However, due to the discrepancy between predictions and explanations in the literature (Rappolt-Schlichtmann et al., 2007; Franse et al., under review) using the explanations to interpret predictions is limited in its validity. Nonetheless, the predictive power of prediction class membership on explanation class membership does suggest that children's predictions about common objects were related to how they went on to explain these objects' buoyancy. As this was not the case for the cubes, this could suggest that the (multiple) features or facts used to discern objects' buoyancies are more easily explicable than those used to predict cubes, perhaps because they are easier to detect or recall. Relatedly, the higher the grade, the more children belonged to the "fact, M, material" class, supporting the idea that many children go on to develop explanations of objects' buoyancies based on multiple sources of information.

What is new in this investigation of buoyancy explanations is the open-ended manner of analysis. Past studies have assigned children to particular explanation types in a hierarchy from least to most advanced (Franse et al., under review) coded the complexity level of children's explanations (Tenenbaum et al., 2004; Rappolt-Schlichtmann et al., 2007) or examined explanation profiles on the basis of answers to multiple choice buoyancy questions designed to reflect previously identified concepts (Schneider and Hardy, 2013; Edelsbrunner et al., 2018). Here coding all of the content of what children said and analyzing this with LCAs meant that we did not need *a priori* hierarchies or expected concepts, allowing us to discover relations in all that children say (see also van der Maas and Straatemeier, 2008). This method, for example, revealed that all explanation classes also mentioned things that fell into the other category, thus indicating that children do not have purely singular explicit concepts of buoyancy. Like any other explanation coding approaches, the findings here are still influenced by how the explanations were coded to begin with (e.g., what falls under other and what receives its own category). It is also important to note the explorative nature of this approach, and future work is needed to corroborate or extend these findings.

Variation and Inquiry-Based Learning

This study indicates that providing children with the opportunity to explore the buoyancy of cubes stimulates the subsequent use of an advanced solution strategy when predicting whether novel cubes will float or sink. This benefit of using an item set that varies only in the relevant features is in line with the suggestions derived from category learning research (e.g., Sloutsky, 2010; Goldwater et al., 2018). In the systematic condition, the exploration cubes in effect presented children with dense categories; the only apparent features making the cubes either float or sink were the masses and volumes (hence, cubes' densities were the defining feature between floaters and sinkers), thereby steering children to hone in on these features. In the non-systematic condition, however, multiple features might have been able to explain why some of the objects float and others sink, making it difficult for the children to identify mass and volume as the relevant features. Although the present findings require replication in future research, the successful findings of this very brief exploration session do hold promise for the use of variation and category denseness to design learning materials.

To understand how, this research can be placed in the context of inquiry-based learning. Across content domains, research indicates that well-guided inquiry-based learning is superior to explicit instruction and also that the presence and the quality of guidance is essential (Alfieri et al., 2011; Lazonder and Harmsen, 2016; Dobber et al., 2017). This is reiterated across several buoyancy studies, indicating that the more guidance the better and more prolonged the learning effects (Hardy et al., 2006; Rappolt-Schlichtmann et al., 2007; Hsin and Wu, 2011). Such guidance can be provided in different ways. For example, the present study used prompts to encourage children to test all of the items and to think about why they float or sink. Relatedly, the floating and the sinking sheets served as status overviews encouraging children to sort items, thus illustrating exploration progress. Perhaps the least explicit type of guidance is process constraints, which restrict the breadth of the learning task (Lazonder and Harmsen, 2016). The simplification of learning materials to vary only in density (i.e., the cubes) could be viewed as a process constraint as this restricted children's exploration to the relevant features.

Whereas several studies have extensively investigated the longitudinal learning effects of mixed-method, long-term buoyancy interventions (Hardy et al., 2006; Schneider and Hardy, 2013; Leuchter et al., 2014; Edelsbrunner et al., 2018; Schalk et al., 2019), the aim of the current experimental study was simply to zoom in on the effect of item variation during hands-on exploration. Nevertheless, by placing the current findings in the context of inquiry-based learning, this study, like those classroom studies, emphasizes the importance of both considering fundamental learning processes in science education design and introducing science early (Leuchter et al., 2014). First, basing the design of inquiry-based learning guidance on a simple model of children's category learning, as was done in this study, seems to provide an effective means of structuring the learning context. In future work, these ideas can be further worked out to provide an informed yet simple blueprint for designing inquiry-based learning founded on the

(category) structure of the to-be learned concepts. This is important as teachers have been shown to not always be fully equipped with the relevant scientific knowledge or insights into children's science learning (Rice, 2005; Leuchter et al., 2014). Second, the findings indicate that even in the first years of primary school, children can (implicitly) apply advanced multidimensional solution strategies when provided with the right environment, thereby echoing calls for starting science education early on using structured hands-on materials and learning environments (Leuchter et al., 2014).

In conclusion, this study investigated the effect of offering items that vary only in the relevant features (i.e., mass and volume) on children's subsequent ability to predict and explain new items' floating and sinking. In line with expectations, children who had had 5 min of hands-on experience with a set of cubes in the water basin were more accurate in predicting the buoyancy of new cubes than children who had used common objects to explore with. Moreover, these children were more likely to use a more advanced solution strategy that decided the buoyancy of an item based on the integration of its mass and volume than the children who had experimented with common objects. These findings suggest that inquiry-based learning designs should consider how to optimize the variation to facilitate learning.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Faculteit der Sociale Wetenschappen Ethiek Commissie (ECPW) Instituut Pedagogische Wetenschappen. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

JS, TS, RF, and MR conceived this project. JS and TS coordinated the data acquisition. JS and MR performed the data analyses. JS wrote the manuscript draft. TS, RF, and MR revised the manuscript draft. All authors contributed to the article and approved the submitted version.

FUNDING

JS, this project, and open access publication was funded by a "Postdocs in Educational Sciences" grant awarded to JS by the Netherlands Initiative for Education Research (NRO; 405-16-381).

ACKNOWLEDGMENTS

We thank the children and their parents, the primary schools, and NEMO Science Museum for their participation and facilitation of this research. We also thank the involved Bachelors and Masters students of Leiden University and the Vrije Universiteit Amsterdam for their assistance with data collection and processing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01665/full#supplementary-material

REFERENCES

Alfieri, L., Brooks, P., Aldrich, N. J., and Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *J. Educ. Psychol.* 103, 1–18. doi: 10.1037/a0021017

Ashby, F. G., and Ell, S. W. (2001). The neurobiology of human category learning. *Trends Cogn. Sci.* 5, 204–210. doi: 10.1016/S1364-6613(00)01624-7

Bouwmeester, S., Sijtsma, K., and Vermunt, J. K. (2004). Latent class regression analysis for describing cognitive developmental phenomena: an application to transitive reasoning. *Eur. J. Dev. Psychol.* 1, 67–86. doi: 10.1080/17405620344000031

Boyer, T. W., Levine, S. C., and Huttenlocher, J. (2008). Development of proportional reasoning: where young children go wrong. *Dev. Psychol.* 44, 1478–1490. doi: 10.1037/a0013110

Butts, D. P., Hofman, H., and Anderson, M. (1993). Is hands-on experience enough? A study of young children's view of sinking and floating objects. *Science* 5, 50–64. doi: 10.1007/bf03170644

Campbell, C. (2015). "Learning theories related to early childhood science education," in *Science in Early Childhood*, eds C. Campbell, W. Jobling, and C. Howitt (Cambridge, MA: Cambridge University Press), 50–64. doi: 10.1017/cbo9781139923804.006

Diamond, A. (2013). Executive functions. *Annu. Rev. Psychol.* 64, 135–168. doi: 10.1146/annurev-psych-113011-143750

Dienes, Z., and Perner, J. (1999). A theory of implicit and explicit knowledge. *Behav. Brain Sci.* 22, 735–755. doi: 10.1017/S0140525X99002186

Dobber, M., Zwart, R., Tanis, M., and van Oers, B. (2017). Literature review: the role of the teacher in inquiry-based education. *Educ. Res. Rev.* 22, 194–214. doi: 10.1016/j.edurev.2017.09.002

Edelsbrunner, P. A., Schalk, L., Schumacher, R., and Stern, E. (2018). Variable control and conceptual change: a large-scale quantitative study in elementary school. *Learn. Individ. Differ.* 66, 38–53. doi: 10.1016/j.lindif.2018.02.003

Erickson, M. A., and Kruschke, J. K. (1998). Rules and exemplars in category learning. *J. Exp. Psychol. Gen.* 127, 107–140. doi: 10.1037/0096-3445.127.2.107

Franse, R. K. (2007). "Science is Primary," in *Handboek Voor Onderzoeken en Ontwerpen in Groep 1En 2*. Amsterdam: Stichting nationaal centrum voor Wetenschap en technologie.

Goldwater, M. B., Don, H. J., Krusche, M. J. F., and Livesey, E. J. (2018). Relational discovery in category learning. *J. Exp. Psychol. Gen.* 147, 1–35. doi: 10.1037/xge0000387

Hardy, I., Jonen, A., Möller, K., and Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students'

understanding of "floating and sinking.". *J. Educ. Psychol.* 98, 307–326. doi: 10.1037/0022-0663.98.2.307

Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., and Trezise, K. (2018). Informative tools for characterizing individual differences in learning: latent class, latent profile, and latent transition analysis. *Learn. Individ. Differ.* 66, 4–15. doi: 10.1016/j.lindif.2017.11.001

Hofman, A. D., Visser, I., Jansen, B. R. J., and Van Der Maas, H. L. J. (2015). The balance-scale task revisited: a comparison of statistical models for rule-based and information-integration theories of proportional reasoning. *PLoS One* 10:0136449. doi: 10.1371/journal.pone.0136449

Hsin, C.-T., and Wu, H.-K. (2011). Using scaffolding strategies to promote young children's scientific understandings of floating and sinking. *J. Sci. Educ. Technol.* 20, 656–666. doi: 10.1007/s10956-011-9310-7

Huang, G.-H., and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika* 69, 5–32. doi: 10.1007/BF02295837

Jones, L. A. (1986). Perception of force and weight: theory and research. *Psychol. Bull.* 100, 29–42. doi: 10.1037/0033-2909.100.1.29

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Kemmers, P., van Graft, M., and Klein Tank, M. (2007). *Drijven en zinken. In Onderzoekend en Ontwerpend Leren bij Natuur en Techniek – Lesmateriaal*. Enschede: Stichting leerplanontwikkeling .

Kloos, H. (2008). "Will it float? How invariance affects children's understanding of object density," in *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, eds B. C. Love, K. McRae, and V. M. Sloutsky Austin, TX: Cognitive Science Society, 687–692.

Kloos, H., and Sloutsky, V. M. (2008). What's behind different kinds of kinds: effects of statistical density on learning and representation of categories. *J. Exp. Psychol. Gen.* 137, 52–72. doi: 10.1037/0096-3445.137.1.52

Kohn, A. S. (1993). Preschoolers' reasoning about density: will it float? *Child Dev.* 64, 1637–1650. doi: 10.1111/j.1467-8624.1993.tb04204.x

Kuhn, D., Iordanou, K., Pease, M., and Wirkala, C. (2008). Beyond control of variables: what needs to develop to achieve skilled scientific thinking? *Cogn. Dev.* 23, 435–451. doi: 10.1016/j.cogdev.2008.09.006

Lazonder, A. W., and Harmsen, R. (2016). Meta-analysis of inquiry-based learning. *Revi. Educ. Res.* 86, 681–718. doi: 10.3102/0034654315627366

Leuchter, M., Saalbach, H., and Hardy, I. (2014). Designing science learning in the first years of schooling. an intervention study with sequenced learning material on the topic of "floating and sinking.". *Int. J. Sci. Educ.* 36, 1751–1771. doi: 10.1080/09500693.2013.878482

McCutcheon, A. (1987). *Latent Class Analysis. Latent Class Analysis. 2455 Teller Road*. Thousand Oaks CA: SAGE Publications, Inc.

Messer, D. J., Pine, K. J., and Butler, C. (2008). Children's behaviour and cognitions across different balance tasks. *Learn. Instr.* 18, 42–53. doi: 10.1016/j.learninstruc.2006.09.008

Mundkur, R. (2015). *Lesson Plan Sink or Float*. Avaliable online at: https://www.education.com/lesson-plan/sink-or-float/ (accessed August 2, 2019).

Muthén, L. K., and Muthén, B. O. (2017). *MPlus user's Guide*, Vol. 8.). Los Angeles, CA: Muthén & Muthén.

Penner, D. E., and Klahr, D. (1996). The interaction of domain-specific knowledge and domain-general discovery strategies: a study with sinking objects. *Child Dev.* 67, 2709–2727. doi: 10.1111/j.1467-8624.1996.tb01884.x

Piaget, J., and Inhelder, B. (1974). *The Child's Construction of Quantities: Conservation and Atomism*. London: Routledge and Kegan Paul.

Pine, K., and Messer, D. (1999). What children do and what children know: looking beyond success using Karmiloff–Smith's RR framework. *New Ideas Psychol.* 17, 17–30. doi: 10.1016/S0732-118X(98)00014-2

Pine, K., and Messer, D. (2003). The development of representations as children learn about balancing. *Br. J. Dev. Psychol.* 21, 285–301. doi: 10.1348/026151003765264093

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rappolt-Schlichtmann, G., Tenenbaum, H. R., Koepke, M. F., and Fischer, K. W. (2007). Transient and robust knowledge: contextual support and the dynamics of children?s reasoning about density. *Mind Brain Educ.* 1, 98–108. doi: 10.1111/j.1751-228x.2007.00010.x

Rice, D. C. (2005). I didn't know oxygen could boil! What preservice and inservice elementary teachers' answers to "simple" science questions reveals about their subject matter knowledge. *Int. J. Sci. Educ.* 27, 1059–1082. doi: 10.1080/09500690500069426

Schalk, L., Edelsbrunner, P. A., Deiglmayr, A., Schumacher, R., and Stern, E. (2019). Improved application of the control-of-variables strategy as a collateral benefit of inquiry-based physics education in elementary school. *Learn. Instr.* 59, 34–45. doi: 10.1016/j.learninstruc.2018.09.006

Schneider, M., and Hardy, I. (2013). Profiles of inconsistent knowledge in children's pathways of conceptual change. *Dev. Psychol.* 49, 1639–1649. doi: 10.1037/a0030976

Science NetLinks (2020). *Sink or Float?*. Avaliable online at: http://sciencenetlinks.com/lessons/sink-or-float/ (accessed August 2, 2019).

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monogr. Soc. Res. Child Dev.* 46, 1–84. doi: 10.2307/1165995

Siegler, R. S., and Chen, Z. (2008). Differentiation and integration: guiding principles for analyzing cognitive change. *Dev. Sci.* 11, 433–448. doi: 10.1111/j.1467-7687.2008.00689.x

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends Cogn. Sci.* 7, 246–251. doi: 10.1016/S1364-6613(03)00109-8

Sloutsky, V. M. (2010). From perceptual categories to concepts: what develops? *Cogn. Sci.* 34, 1244–1286. doi: 10.1111/j.1551-6709.2010.01129.x

Smith, C., Carey, S., and Wiser, M. (1985). On differentiation: a case study of the development of the concepts of size, weight, and density. *Cognition* 21, 177–237. doi: 10.1016/0010-0277(85)90025-3

Tenenbaum, H. R., Rappolt-Schlichtmann, G., and Zanger, V. V. (2004). Children's learning about water in a museum and in the classroom. *Early Child. Res. Q.* 19, 40–58. doi: 10.1016/j.ecresq.2004.01.008

van der Maas, H. L. J., and Straatemeier, M. (2008). How to detect cognitive strategies: commentary on "Differentiation and integration: guiding principles for analyzing cognitive change.". *Dev. Sci.* 11, 449–453. doi: 10.1111/j.1467-7687.2008.00690.x

Visser, I., and Speekenbrink, M. (2010). depmixS4: an R Package for hidden markov models. *J. Stat. Softw.* 36, 1–21.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 van Schaik, Slim, Franse and Raijmakers. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 22 July 2020
doi: 10.3389/fpsyg.2020.01621

Inquiry-Based Learning and Conceptual Change in Balance Beam Understanding

*Joep van der Graaf**

Behavioural Science Institute, Radboud University, Nijmegen, Netherlands

Inquiry-based learning has the potential to foster conceptual change, but whether it can induce an advancement in strategy use is not yet known. Such an advancement seems plausible, because conceptual change can be reflected in the use of new strategies. Whether inquiry-based learning leads to advancement in strategy use can be tested with strategy-based tests, such as the balance beam test. Distinct strategies have been proposed and identified for this test. Therefore, the present study compared response patterns on the balance beam test before and after an inquiry-based lesson. The experimental condition completed a digital inquiry-based lesson about the balance beam (n = 113), and the control condition completed a similar inquiry-based lesson but investigated a different topic (n = 44). The participants were aged 8–13 years old and were unfamiliar with the law of moments. The balance beam test (pretest and posttest) consisted of 25 items. Overall accuracy in solving balance beam problems improved after the inquiry-based lesson in the balance beam (BB) condition but not in the control condition. Classes, identified with latent transition analysis (LTA), appeared to be globally in line with previously identified strategies in the balance beam test. Condition was entered as a covariate in the LTA to identify which changes in strategy use could be attributed to the experimental intervention. First, changes from pretest to posttest were found, which supported that a change in strategy use occurred in some children. Second, there were more improvements in the BB condition, and these improvements indicated larger gains compared to the control condition. This means that in science education, it is important to map out prior knowledge and its effect on learning paths. Overall, results suggested that conceptual change could be measured as a change in strategy use and modeled with LTA to reveal that 26% of the children showed conceptual change after a single inquiry-based lesson.

Keywords: conceptual change, scientific reasoning, inquiry-based learning, latent transition analysis, strategy use

OPEN ACCESS

Edited by:
Amanda C. Brandone,
Lehigh University, United States

Reviewed by:
Florencia K. Anggoro,
College of the Holy Cross,
United States
Jan Boom,
Utrecht University, Netherlands

***Correspondence:**
Joep van der Graaf
j.vandergraaf@pwo.ru.nl

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *31 January 2020*
Accepted: *16 June 2020*
Published: *22 July 2020*

Citation:
van der Graaf J (2020) Inquiry-Based Learning and Conceptual Change in Balance Beam Understanding. Front. Psychol. 11:1621. doi: 10.3389/fpsyg.2020.01621

INTRODUCTION

Inquiry-based learning entails discovering scientific laws through the process of investigation (Dunbar and Klahr, 2012). It is relevant for contemporary education in which both knowledge and skills are taught (Fischer et al., 2014). Inquiry-based learning is useful in this regard, as it promotes the acquisition of scientific reasoning skills and scientific knowledge

(Lazonder and Harmsen, 2016). During the development of scientific reasoning skills and construction of scientific knowledge, conceptual change takes place. Conceptual change refers to children's gradual revision of beliefs from intuitive (and often incorrect) to complete and correct conceptions; these conceptions, either intuitive or complete, are called explanatory frameworks (Vosniadou, 2002). Explanatory frameworks can be measured as strategy use when solving problems, because strategies reflect children's explanatory frameworks. Previous studies have revealed that, especially, young children (i.e., primary school children) are susceptible to conceptual change (Vosniadou, 2002) and that inquiry-based learning can foster conceptual change (Huang et al., 2017). However, little is known about the effects of inquiry-based learning on strategy use. Therefore, the present study investigated strategy use before and after a digital inquiry-based lesson in primary education. In the present study, conceptual change was measured as a change in strategy use when solving balance beam problems. To foster conceptual change, an inquiry-based lesson about the balance beam was used and compared to a control condition where participants completed an inquiry-based lesson about a different topic, namely, gears.

Conceptual change theories describe how people mentally store and organize information and how this information and its organization change when people learn. There are debates about what a concept is and how change takes place (e.g., Piaget, 1978; Posner et al., 1982; Chi and Roscoe, 2002; diSessa, 2002; Vosniadou, 2002). Concerning the question of what a concept might be, Vosniadou (2002) definition will be used, as it deals with young children's science learning. Conceptual change assumes that there is an initial mental model, or explanatory framework about scientific phenomena (Vosniadou, 2002). These explanatory frameworks embody causal notions and allow distinct types of explanations and predictions (Carey, 1985; as cited in Vosniadou, 2002). Explanations and predictions are constrained by young children's conceptual system, which drives the interpretation of their observations and the information they receive.

For example, when considering the law of moments as a scientific phenomenon that explains which side of a balance beam goes down when placing weights on it, then an example of an explanatory framework is the predictions about which side of a balance beam goes down and the explanations why. A balance beam (or seesaw) has two sides on which different amount of weights can be placed at different distances from the center (i.e., fulcrum). The simplest explanation proposed by Siegler (1976) is that only weight on either side of the balance beam determines which sides goes down (Strategy 1), while the most complex (and the only correct) explanation is that weight and distance on either side should be multiplied and compared to determine which side goes down (Strategy 4). In between are Strategy 2, which is like Strategy 1, but distance is also used when weights are equal, and Strategy 3, using both weight and distance but in an inconsistent manner.

Explanatory frameworks such as these guide learners' behavior in science tasks. Therefore, explanatory frameworks can be measured through a learners' strategy use, and conceptual change can be seen as a change in strategy use. For conceptual change to occur, children can advance from simple strategies that they need to abandon to be able to adopt new (and potentially more complex) strategies. Note that this assumes that initial strategies are error prone and/or incomplete, which is often the case in primary school children (Vosniadou, 2002).

This brings about the question about how change takes place. Conceptual change takes place in roughly two ways: explanatory frameworks themselves change, or the relations among explanatory frameworks change (see diSessa and Sherin, 1998, for a review). For example, in the first case, distance can be added as a factor to the explanatory framework of the law of moments. In the latter case, how weight and distance are combined to determine which side of a balance beam goes down might change from "add weight and distance on either side and then compare the sides" (Addition Strategy) to "multiply weight and distance on either side and then compare the sides" (Strategy 4). There is evidence from studies on young children's understanding of force, matter, heat, etc. that these changes occur slowly (Vosniadou, 2002). Chinn and Brewer (1998) further specify how explanatory frameworks transform describing the degree of change from zero change to complete change: children can ignore data, reject data, question data's validity, exclude data from their respective theory, suspend data (abeyance), reinterpret data, accept data and make peripheral changes to the current theory, or accept data and change theories. To summarize, conceptual change is conceptualized as a change in explanatory frameworks themselves and/or a change in the relations between explanatory frameworks, and both types of changes can range from no change to full change.

The current study explored how inquiry-based learning can promote conceptual change. Although explanatory frameworks might not easily change and conceptual change seems to be a slow process (Vosniadou, 2002), inquiry-based learning has several features that help to promote conceptual change. Inquiry-based learning enables children to gain insight into scientific concepts by engaging in scientific reasoning activities, such as generating hypotheses, conducting experiments, and drawing conclusions (Klahr, 2000). During the first step, generating hypotheses, children think about and explicate what might explain a specific scientific phenomenon, such as how weight might affect which side of the balance beam goes down. In the second step, experimentation, children set up an experiment by selecting settings of variables, such as selecting a weight of 5 kg to be placed on the left side of a balance beam, and observe what happens: left side goes down, right side goes down, or the beam is balanced. The final step, drawing conclusions, is when children interpret what they have observed and explain it, such as that the left side of the beam went down because there was a weight on the left side and no weight on the right side. These scientific reasoning activities should be well guided in order to be effective in enhancing learning (see Lazonder and Harmsen, 2016 for a meta-analysis).

Several processes of inquiry-based learning seem to foster conceptual change. Conceptual change is hampered by children's lack of orderly organization when explaining and predicting scientific phenomena, bias in evaluating evidence, and problems

falsifying or verifying understanding (Vosniadou, 2002). The first problem can be addressed by inquiry-based learning, as children should articulate their explanations and predictions when generating hypotheses, which is one of the first steps in the inquiry cycle. Secondly, self-generated evidence might help to overcome bias in evaluating it (Mussweiler and Strack, 1999), and perceiving evidence can help in creating a new explanation, instead of distorting the evidence (Koerber et al., 2017). Third, inquiry-based learning revolves around proposing explanations and testing them in experiments (Dunbar and Klahr, 2012), which is a process of explicating understanding that subsequently might be falsified or verified. Taken together these features of inquiry-based learning have the potential to induce conceptual change.

There is empirical support that guided inquiry can foster conceptual change. Huang et al. (2017) designed an eight-lesson inquiry-based unit for 13- to 14-year olds that comprised six investigations. During the eight lessons, students were introduced to the topic: motion and force. Each investigation was about a different specific topic, such as changing speed. During investigations, students predicted what the effect of a variable would be (i.e., hypothesis generation), observed what the effect was by designing and conducting experiments, and explained what the effect was (i.e., conclusion). Conceptual change was operationalized as a decrease in misconceptions, an increase in correct propositions, and a correct organization of propositions. Misconceptions were assessed by a multiple-choice test, and the propositions and their organization through a concept map. The results showed that children held fewer misconceptions and more correct propositions after the inquiry-based lesson series—although no improvement was observed in the organization of propositions. Despite this clear advancement of understanding, it remains unclear whether the guided inquiry also led to usage of different explanatory frameworks. A strategy-based assessment of conceptual change is needed.

In addition to evidence about which features of inquiry-based learning might foster conceptual change, there is evidence about how inquiry-based learning might foster conceptual change. In the case of the balance beam task (Siegler and Chen, 1998), through inquiry-based learning, children should notice potential explanatory variables, formulate new explanations, generalize to new problems, and maintain the new strategy to learn new strategies. Inquiry-based learning revolves around these processes. Initial strategies, for example, Strategy 1 (use weight only), can be put to the test. When the results do not match predictions, for example, when the side with the largest weight does not go down, children might notice the effect of distance. Noticing potential explanatory variables seems to be the most important factor in learning new strategies (Siegler and Chen, 1998). The results from these experiments can be used to formulate new predictions. Subsequently, new predictions, such as those in line with Strategy 2 (use distance when weights are equal), can be generalized and tested. The results then might show support for the new predictions. When children explain these findings, they might (partially) adopt and maintain a new strategy, Strategy 2 in this case.

There is large variation within and between children in which strategies are used at a given time and how the strategies that are used change. Therefore, in order to model conceptual change, this natural variation within and between children should be taken into account by appropriate statistical models. Latent models, in particular, are appropriate for two main reasons. Firstly, explanatory frameworks show qualitative and quantitative differences between children, and secondly, explanatory frameworks show variation at specific time points, as well as over time (see Hickendorff et al., 2018, for details). Therefore, statistical models should allow intra- and inter-individual differences at a single time point but also between time points. Models with these features have been applied to strategy-based tests that can measure explanatory frameworks in use. A common example is the balance beam test (Inhelder and Piaget, 1958; Siegler, 1976), which lends itself well to latent class analysis (LCA) (Jansen and van der Maas, 1997). LCA identifies a number of homogeneous subgroups, called classes, in the heterogeneous data of the whole group at a specific point in time and has frequently been used to model strategy use when solving balance beam problems (Jansen and van der Maas, 1997, 2002; Boom et al., 2001; Boom and Ter Laak, 2007). Two common findings are that (1) five or six classes are found and (2) Strategies 1, 2, and 4 are identified as separate classes. Strategy 3 often is not found as such, but a variant can be detected. Instead of complete uncertainty when weight and distance contradict which side goes down (i.e., Strategy 3), children may use the so-called Addition Strategy (Wilkening and Anderson, 1982): weight and distance on both sides are added and then compared (e.g., Boom et al., 2001; Boom and Ter Laak, 2007). The study by Boom et al. (2001) investigated strategy use of 7- to 14-year-old children. In addition to Strategies 1, 2, 4, and the Addition Strategy, two classes (out of six) did not match any strategy. A later study with 5- to 14-year-old participants and two covariates of class membership (age and working memory) revealed comparable classes, but the class representing Strategy 4 was not found (Boom and Ter Laak, 2007). Boom and Ter Laak (2007) attributed the absence of Strategy 4 to the inclusion of only few older participants. These studies show the feasibility of using LCA to identify strategy use and thus capture children's explanatory frameworks.

Latent transition analysis (LTA) models conceptual change by extending LCA to multiple time points. LTA creates classes, as LCA does, but it also creates transition probabilities that indicate the likelihood of moving from one class, at a point in time, to another, at a later point in time (Bartolucci et al., 2014). LTA has previously been applied to inquiry-based learning settings. Schneider and Hardy (2013) studied the effects of a constructivist intervention about floating and sinking using three conditions: baseline, low instructional support, and high instructional support. Participants were children from 8 to 11 years old. Five profiles were found ranging from misconceptions to scientific conceptions. From pretest to posttest (1 week after the intervention), most children improved their understanding, and from posttest to retention test (1 year later), most children did not change. In addition, more instructional support was related to more transitions, indicating improvement. Another example is a study on change in knowledge structures after 15

lessons, which included inquiry-based learning (Edelsbrunner et al., 2018). The authors analyzed primary school children's (ages 6–13 years) knowledge about floating and sinking. Medium-sized effects were found that showed a decrease in misconceptions and an increase in correct conceptions. LTA showed that the largest change was a transition from moderate misconceptions to low misconceptions, indicating a restructuring of children's knowledge. These results are promising in light of studying conceptual change, as LTA was more informative than regular regression-based analyses, which do not take into account differences in students' knowledge and its development (Edelsbrunner et al., 2018).

To summarize, conceptual change can be observed in a change in strategy use, but this notion has received little attention in empirical research. Furthermore, it is yet unknown what the effect of inquiry-based learning on advancement in strategy use is, while it can be disentangled using informative statistical models (Hickendorff et al., 2018). Therefore, the present study investigated strategy use when solving balance beam problems, before and after a digital inquiry-based lesson with a virtual lab about balance beams (BB condition) compared to a control condition. In both conditions, children completed the balance beam test before and after an inquiry-based lesson, but the control condition studied a different topic, namely, gears. Thus, it was the aim to detect change in strategy use when solving balance beam problems after an inquiry-based lesson about the balance beam, compared to a control group that also completed an inquiry-based lesson but on a different topic. The research questions were: (1) to what extent does inquiry-based learning about balance beams (versus inquiry-based learning about gears) promote advancement in strategy use when solving balance beam problems, and (2) to what extent can change in strategy use be measured with the balance beam test and modeled with LTA? In addition, effects of age and gender on overall accuracy at pretest and posttest, and on the difference score (posttest minus pretest), were explored.

Regarding strategy use, it was hypothesized that five or six classes would be found and that Strategies 1, 2, and 4 and the Addition Strategy would be detected, as well as one or two classes that could not be classified, because that is what previous studies with LCA of balance beam performance showed (Jansen and van der Maas, 1997, 2002; Boom et al., 2001; Boom and Ter Laak, 2007). It was expected that accuracy in the present study would be lower than in the study by Boom et al. (2001) because they used four age groups to estimate the classes and the present study targeted a single, younger age group, 10 years and 9 months on average.

With reference to conceptual change, the inquiry-based lesson about the balance beam was hypothesized to be effective in teaching strategies that are more advanced to children. An improvement in strategy use from pretest to posttest was expected in the BB condition, while no improvement was expected in the control condition. In line with these expectations, class membership was expected to change from pretest to posttest in the BB condition and not in the control condition. To assess change of class, LTA was used. It was hypothesized that children would advance in the BB condition from classes matching a simple strategy to classes matching the next strategy, indicating slow advancement (Vosniadou, 2002).

MATERIALS AND METHODS

Participants

A total of 157 children from grades 4, 5, and 6 participated. The BB condition included 114 children, and the control (gears) condition, 45 children. Age ranged from 8 years and 5 months to 13 years (M = 10 years and 9 months, SD = 1 year). The children had not learned about the law of moments or gears yet. Children were not familiar with digital inquiry-based learning, but they had experience with digital learning materials and with inquiry-based learning without digital media.

Participants came from five schools and eight classrooms. Schools agreed to participate in the present study and helped to inform the parents and/or caretakers. Two schools chose to ask for consent actively, and three chose the passive approach. Consent was obtained from the parents and/or caretakers. Children and their parents and/or caretakers were given information about the present study and were given the opportunity to ask questions.

Design

Classrooms were assigned randomly to one of the conditions, while making sure that more classrooms would be assigned to the balancing beam condition to obtain a larger sample size for transitions in the LTA. The procedure was the same in both conditions: There were three sessions; each on a separate day with 1 week between them. In the first session (pretest), the balance beam test was administered (duration was 15 min). The second session consisted of the digital inquiry-based lesson (duration was 45 min). The third session (posttest) consisted of the balance beam test. Children also completed other tests during the first and third sessions, as part of a larger study.

All tests and the digital inquiry-based lesson were performed on a computer. Children went to the computer room, class by class, or laptops were used in the classroom. The instruction was presented digitally. Before children started, the experimenters instructed them to log in and proceed with the instruction and tests. The teacher was present during all sessions to manage the classroom.

Materials

Inquiry-Based Learning Environment

Children in both conditions completed a digital inquiry-based learning lesson. The BB condition studied the effects of weight and distance on the balance of balance beams. The control condition investigated gears, their direction of rotation when creating chains of at least two gears, and acceleration when connecting different sizes of gears. The lessons were similar in all aspects (introduction, increase in difficulty, sequence of activities, and duration) except for the topic of investigation. Thus, both lessons were comparable in rigor and complexity. Children worked at their own pace. There were no practice trials at the beginning, but there was instruction on how to use the

digital learning environment. An experimenter was present to answer questions regarding the use of the learning environment.

In the introduction in the BB condition, a problem was presented in which a seesaw did not seem to work because the mother was heavier than the child was. It was suggested that asking a friend to join (i.e., change the weight) or asking the mother to move closer to the fulcrum (i.e., change the distance) might solve the problem, with pictures of the problem and both suggestions. Next, in the experimentation steps, children were instructed to manipulate one variable (weight or distance) and keep the other one constant (distance or weight). This was done first for weight and then for distance. For the effect of both, children were instructed to manipulate both weight and distance. Furthermore, in all experimentation steps, children could type information about the experimental design and the outcomes in an input box. The last part of the instruction in the experimentation phase was that after several tries, they were asked to explain, as they would to someone who did not understand it yet, how they managed to balance the beam and what the effect of weight, distance, or both was on balance.

Experiments were conducted by manipulating weight and distance in a virtual lab; see **Figure 1**. The lab directly showed the effect of the manipulations on the balance of the beam, which provided feedback to the children about the role of the variable(s) they manipulated. The brief explanation in the final step included the notion that both weight and distance matter for the balance of a beam, and there was a short quiz, so children could use their newly acquired knowledge. The short quiz consisted of two multiple-choice questions with four options about where to place a person on one side of a seesaw to make sure that it would be balanced. There was feedback on the correctness of the response and an explanation about the correct option. A hint was available, and when clicked, it showed an example in which torque was calculated. Finally, children were thanked for their participation.

In the control condition, the introduction presented a problem of a boy who wanted to cycle backward and do it fast. In the experimentation steps, the number of gears and direction of rotation, the type of connection between gears (direct or with a belt) and direction of rotation, and different sizes of gears and acceleration were investigated. The lesson ended with more information about gears, direction of rotation, and acceleration.

Balance Beam Test

The balance beam test aimed to assess strategy use when solving problems with the balance beam (Inhelder and Piaget, 1958; Siegler, 1976). Children were asked to decide which side of the balance beam would go down, left, right, or neither (balance). The weights and distances as described by Boom et al. (2001) were used in the same order to create 25 items divided into five sets, consisting of five items each.[1] The first set of five items addressed weight. The distances were the same on each side of the beam, but the weights were not. The second set of five items addressed distance: weights were the same, but distances were not. The third set of five items addressed balance: weights and distances were different but in such a way that the beam would balance. The fourth set of five items addressed conflict-weight, which means that the larger weight goes down, but distance on that side is smaller than on the other side. The last set of five items addressed conflict-distance, in which the side with the larger distance does go down, but weight on that side is smaller. The weights and distances in conflict-weight and conflict-distance items were different for each side of the balance beam. A digital version of the test was created for the present study. Screenshots with various setups of a balance beam in a virtual lab, called Balancing Act, were used; see **Figure 2**. Balancing Act was produced by the PhET project at the University of Colorado, Boulder[2]. At the bottom of the screen, a calculator could be accessed. If children wanted, they could access it to do calculations needed to solve the problems during the test. When using Strategy 1 (weight only), the weight and conflict-weight items should be answered correctly, and the rest incorrectly. Strategy 2 (use weight and, when weights are equal, use distance) should lead to correct responses on weight, conflict-weight, and distance items but incorrect responses on conflict-distance and conflict-balance items. When applying Strategy 3 (using both weight and distance but in an inconsistent manner), weight and distance items should be answered correctly, and all other items should have accuracy around chance (33% correct). The Addition Strategy (add weight and distance) would lead to correct answers on weight, distance, and conflict-balance items, as well as three of the conflict-weight items and one conflict-distance item; see **Appendix A**. Finally, when all items are answered correctly, accuracy indicates usage of Strategy 4 (use weight and distance correctly). The test consisted of 25 items in total, and for each item, one point could be earned. The test was reliable (Kline, 1993) at pretest, $\alpha = 0.72$, $\lambda_2 = 0.77$, and posttest, $\alpha = 0.73$, $\lambda_2 = 0.77$.

Analyses

The analysis of balance beam performance on pretest and posttest mainly followed the analyses by Boom et al. (2001). The number of test items answered correctly was calculated for each child. One control participant had more than six unanswered items on pretest and was removed from further analyses. All other participants did answer at least 19 items. Unanswered items were recoded as incorrect, because missing per variable was low (maximum 4.5%) and visual inspection revealed no common pattern of missingness.

LTA was used to assign response patterns on all items to latent classes. The classes were based on both pretest and posttest; thus, each participant contributed two sets of responses in the LTA. Classes were categorized and named based on their resemblance to Siegler's strategies (1976). In addition, transition probabilities were calculated that indicate the likelihood of moving from one class at pretest to another at posttest. Condition was added as a covariate of transition probabilities. To match the analysis by Boom et al. (2001), who used LCA, a latent Markov (LM) model

[1]Three items deviated from the study by Boom et al. (2001), due to incorrect digitization. Two items did not change in type, but one item did, namely, item 25 about conflict-distance. It was therefore a conflict-weight item; see **Appendix A** for all item details. Two items were identical, namely, items 16 and 25. Since accuracy was highly similar and an additional item provided more information for the analysis, item 25 was not removed.

[2]https://phet.colorado.edu

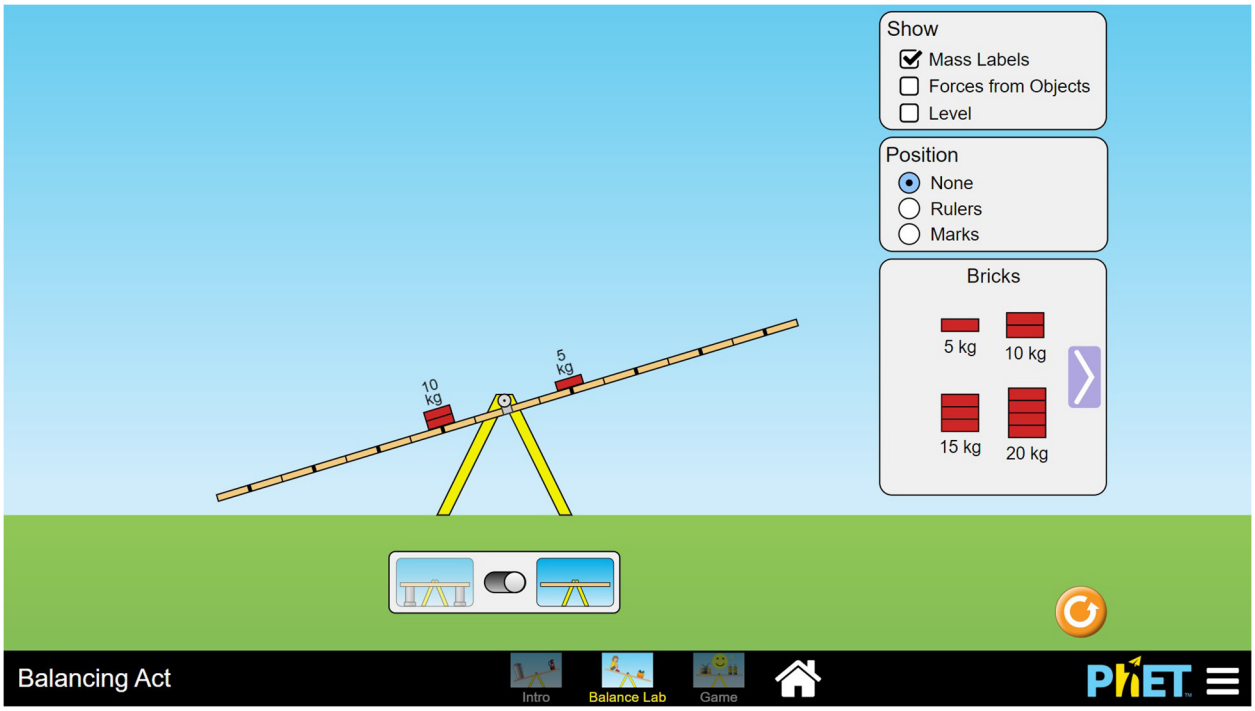


FIGURE 1 | The virtual lab with the balance beam that was used during the inquiry-based lesson.

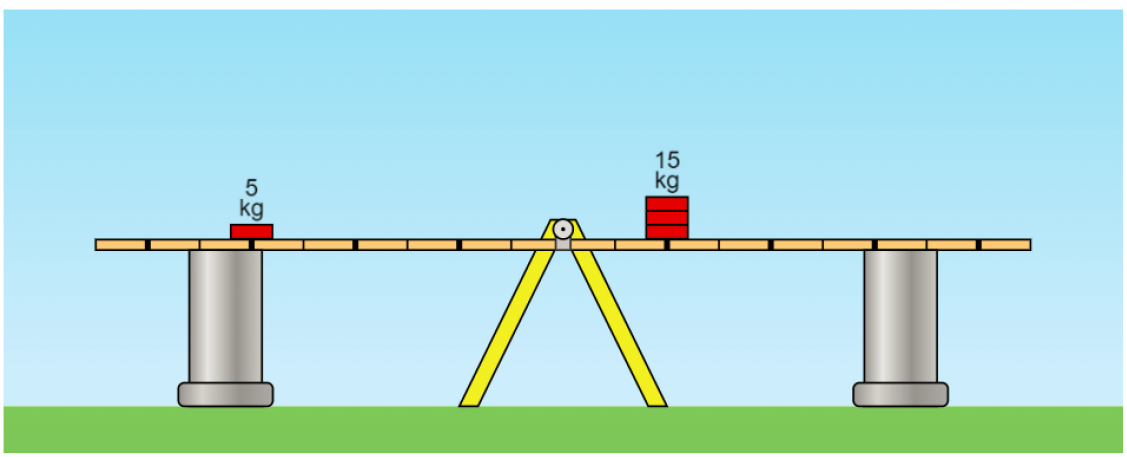


FIGURE 2 | An item of the balance beam test.

for longitudinal data was used, which has the same features, namely, Bayesian inferencing and Gibbs sampling (Bartolucci et al., 2014). This is also in line with the LTA by Edelsbrunner et al. (2018). The LMest package (Bartolucci et al., 2017) was used to conduct LTA with the present data using LM models. This package is available in R (R Core Team, 2019). Transition probabilities were set to be dependent on time, which means that transitions from pretest to posttest only depended on class assignment at pretest. The tolerance level was set to $10e^{-8}$. The maximum number of iterations of the algorithm was set to 1,000. This resulted in conditional response probabilities per item for each class, also called item probabilities, which indicate the probability to answer the item correctly. In addition, transition probabilities were calculated for each class on pretest to each class at posttest. Finally, for each participant, the probability to belong to a class was calculated per class. Note that one participant can be assigned to multiple classes, but often, one class is dominant (over 75% assignment). This analysis takes into account effects per time point and over time when calculating class assignment and transition probabilities (Bartolucci et al., 2017).

RESULTS

Comparing the Balance Beam and Control Condition

The first set of analyses examined whether participants in the BB and control conditions differed in age, gender, and test performance (pretest, posttest, and difference score); see **Table 1**. Participants in the control condition were slightly younger than in the BB condition. Other comparisons between the conditions revealed no difference in gender or pretest scores but a higher posttest and difference score (posttest minus pretest) for the BB condition.

Age and Gender Effects on Balance Beam Performance

The second set of analyses examined how age and gender were related to test performance. Correlations were calculated first. Age and the difference score correlated (see **Table 2**), which can be explained by younger participants on average in the BB condition. Pretest, posttest, and difference scores correlated as expected.

To further explore age and gender effects, a set of regression analyses were conducted. Given the sample size, using age and gender as covariates in the LTA was not feasible. The regressions had condition, age, gender, and all their interactions as independent variables that were related to total correct on pretest and posttest and the difference score as dependent variables. Both forward and backward regression were conducted, and the results were the same. Regarding the total correct on pretest, there was an interaction effect of age and condition, $\beta = 0.17$, $t(144) = 2.02$, $p = 0.045$. Follow-up analyses showed that total correct on pretest and age were correlated in the BB condition, $r(101) = 0.23$, $p = 0.020$, but not in the control condition, $r(43) = -0.020$, $p = 0.898$. This indicated that in the BB condition, which included more and younger children on average, age and balance beam understanding were positively related.

With reference to the total correct on posttest, only condition showed a significant relation, $\beta = 0.39$, $t(144) = 5.08$, $p < 0.001$. Total correct on posttest was larger in the BB compared to the control condition; see **Table 1**.

Concerning the difference score, the effect of condition, $\beta = 0.68$, $t(143) = 3.11$, $p = 0.002$, and the interaction of condition with age were significant, $\beta = -0.44$, $t(143) = 2.01$, $p = 0.047$. Difference scores were larger in the BB compared to the control condition; see **Table 1**. Follow-up analyses showed that the difference score and age were correlated in the BB condition, $r(101) = -0.21$, $p = 0.031$, but not in the control condition, $r(43) = -0.020$, $p = 0.902$. This indicated that in the condition that showed learning, the younger the child, the more he or she learned, on average.

Learning Gains

To verify that learning occurred, pretest and posttest scores were analyzed while taking condition into account in a mixed-design ANOVA. Condition (control vs. BB) was the between-subjects factor, and time (pretest vs. posttest) the within-subjects factor. First, age was added as a covariate to control for possible age effects. The interaction of age with time and the three-way interaction of age, condition, and time were not significant, $F < 1$, $p > 0.350$ in all cases. Therefore, the repeated-measures ANOVA was conducted without age. The main effect of time was not significant, $F(1,155) = 0.97$, $p = 0.327$, $\eta_p^2 = 0.006$, but the interaction of time and condition was significant, $F(1,155) = 13.36$, $p < 0.001$, $\eta_p^2 = 0.079$. This interaction effect can be explained by higher posttest scores in the BB condition, while pretest scores did not significantly differ from the control condition; see **Table 1**. To further explain this result, pretest and posttest scores were compared, and posttest scores were higher than pretest scores in the BB condition, $t(112) = 4.72$, $p < 0.001$, $d = 0.38$, but not in the control condition, $t(43) = 1.35$, $p = 0.184$, $d = 0.22$. The effect in the BB condition was small (Cohen, 1992).

Latent Transition Analysis

Before analyzing change in strategy use, the number of classes was determined. This was determined based on both pretest and posttest performance. LTA was conducted with condition as a covariate of transitions from pretest to posttest because condition affected learning gain. To verify that condition did not affect class assignment on pretest, LTA was conducted with condition as a covariate of initial probabilities as well; condition did not affect these probabilities. In other words, for any class at pretest, children were as likely to be assigned to it in the BB as in the control condition. Thus, LTA was conducted with condition as

TABLE 1 | Descriptive statistics and comparison of the balance beam (BB) and control conditions on gender, age, pretest and posttest balance beam scores, and the difference between pretest and posttest.

	BB			Control						
	M	*SD*	*N*	*M*	*SD*	*N*	*df*[1]	*t*	Adj. *p*	*d*
Gender[2]	1.48	0.50	105	1.41	0.50	44	147	0.75	0.457	0.13
Age*	10.60	1.08	102	11.00	0.69	44	122.61	2.67	0.014	0.41
Pretest	13.90	3.86	113	13.36	4.02	44	155	0.80	0.457	0.14
Posttest*	15.28	3.34	113	12.57	3.78	44	155	4.41	<0.001	0.78
Difference score*	1.38	3.11	113	−0.80	3.91	44	155	3.52	<0.001	0.65

[1]*When variances were unequal, degrees of freedom of the t-tests were adjusted.* [2]*Girl = 1, boy = 2.* *$p < 0.05$.

TABLE 2 | Pearson's correlations for condition, gender, age, pretest and posttest balance beam scores, and the difference between pretest and posttest.

	1	2	3	4	5	6
(1) Condition	–					
(2) Gender	0.06	–				
(3) Age	−0.18*	−0.10	–			
(4) Pretest	0.06	−0.04	0.15	–		
(5) Posttest	0.33**	0.09	−0.05	0.56**	–	
(6) Difference score	0.28**	0.13	−0.20*	−0.49**	0.45**	–

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

TABLE 3 | Fit of the models with two to seven classes.

Number of classes	Number of free parameters	Log-likelihood	AIC	BIC
2	55	−3,787.38	7,684.75	7,851.43
3	89	−3,499.31	7,176.63	7,446.33
4	127	−3,478.22	7,210.45	7,595.31
5	169	−3,195.29	6,728.57	7,240.71
6	215	−3,153.34	6,737.80	7,389.34
7	265	−3,071.76	6,673.51	7,476.58

AIC stands for Akaike information criterion, and BIC stands for Bayesian information criterion.

a covariate of transitions only. To determine the number of classes, the fit of models with two to seven classes was investigated using random initialization of the expectation-maximization algorithm, and when the optimal number of classes was found, a deterministic initialization was used (Bartolucci et al., 2017). In a first analysis, two classes appeared to be a result of outliers, as these classes were on either end of the scale (very low accuracy and very high accuracy). Furthermore, these classes were barely used (below 2%). Therefore, these four outliers (all in the BB condition) were removed, and LTA was conducted with a sample size of 153. The Bayesian information criterion (BIC), which is a good indicator of fit in models with latent variables (Nyland et al., 2007), supported a five-class model (i.e., the lowest BIC); see **Table 3**. Using different starting values did not make a difference; the five-class solution had the lowest BIC. An entropy of 0.89 for this solution indicated that separation between the latent classes was good, as it approached 1 (Celeux and Soromenho, 1996).

The Classes: Modeling Strategy Use

Response patterns (based on the item probabilities) of the five classes were investigated in order to characterize the classes; see **Figure 3**. **Figure 3** depicts how children in the different classes scored on all items of the balance beam test. Class D was used most often. Class E had the highest overall accuracy on the balance beam test; see **Table 4**.

Class A did not match any of the proposed strategies, and it was called "Unstructured." The pattern did not reveal consistency in the responses. Accuracy on all items was around chance (0.33 item probability). The only exception was weight, where performance was slightly better. Class A, however, does not match Strategy 1 (use weight only in predicting which side goes down), because accuracy on conflict-weight items is around chance. This is interesting to note, because such patterns would not be visible when excluding the weight-only items (cf. Boom et al., 2001). Class A was used 14.19% of the time on pretest.

Class B had a good fit with Strategy 1, using weight only and ignoring distance, and was, therefore, called "Strategy 1." This can be seen by the high accuracy on weight and conflict-weight items, while accuracy on other items is around zero. This large difference in accuracy between weight and non-weight items is characteristic of Strategy 1. Class B was used 15.25% of the time on pretest.

Class C did not match any of the strategies proposed by Siegler (1976) and was called "Unclassified." Class C could be indicative of Strategy 3 to some extent: using both weight and distance but in an inconsistent manner. Class C showed a large resemblance to class 3 found by Boom et al. (2001). The item probabilities were large for weight and distance items, indicating that most participants in this class were able to solve these problems. The confidence intervals were larger for the conflict items than for non-conflict items, indicating that there was a larger variability of responses on conflict items. Conflict-distance item 1 had a relatively high accuracy compared to the other conflict-distance items, which might be explained by the relatively high torque of this item; see **Appendix A** for item details. Class C was used 21.91% of the time on pretest.

Class D had a good fit with Strategy 2 (i.e., use weight and, when weights are equal, use distance) and was called "Strategy 2." The item probabilities indicated high accuracy on weight and conflict-weight items. Accuracy on distance items was high and close to 100% for three items. The overall pattern of class D showed a good match with class 2 from Boom et al. (2001). Class D was used 27.31% of the time on pretest.

Class E had a reasonable fit with the Addition Strategy. When using the Addition Strategy, weight and distance are added on each side, and subsequently, these sums are compared. All conflict-balance items; conflict-weight items 3, 4, and 5; and conflict-distance item 1 should be answered correctly when using the Addition Strategy. Accuracy on these items in class E was around 75% correct. Boom et al. (2001) also found a class matching the Addition Strategy: class 4. Class E was used 21.31% of the time on pretest.

To summarize, classes B and D were shown to be a good match to Strategies 1 and 2 proposed by Siegler (1976). Class E matched the Addition Strategy (Wilkening and Anderson, 1982). Class C did not match any strategy, but it is in line with previous studies that found such classes (e.g., Boom et al., 2001). Class A showed an unstructured pattern and did not match any previously detected strategy.

Each child was assigned a probability to belong to the classes. There were 112 children who were assigned to a single class (over 99% assignment). Additionally, 35 children had a 75% or higher assignment to a single class. Only six children had a lower than 75% probability to belong to one class, but the lowest probability for these children was over 50%. At posttest, these numbers were comparable: 113 children over 99%, 32 children over 75%, 7

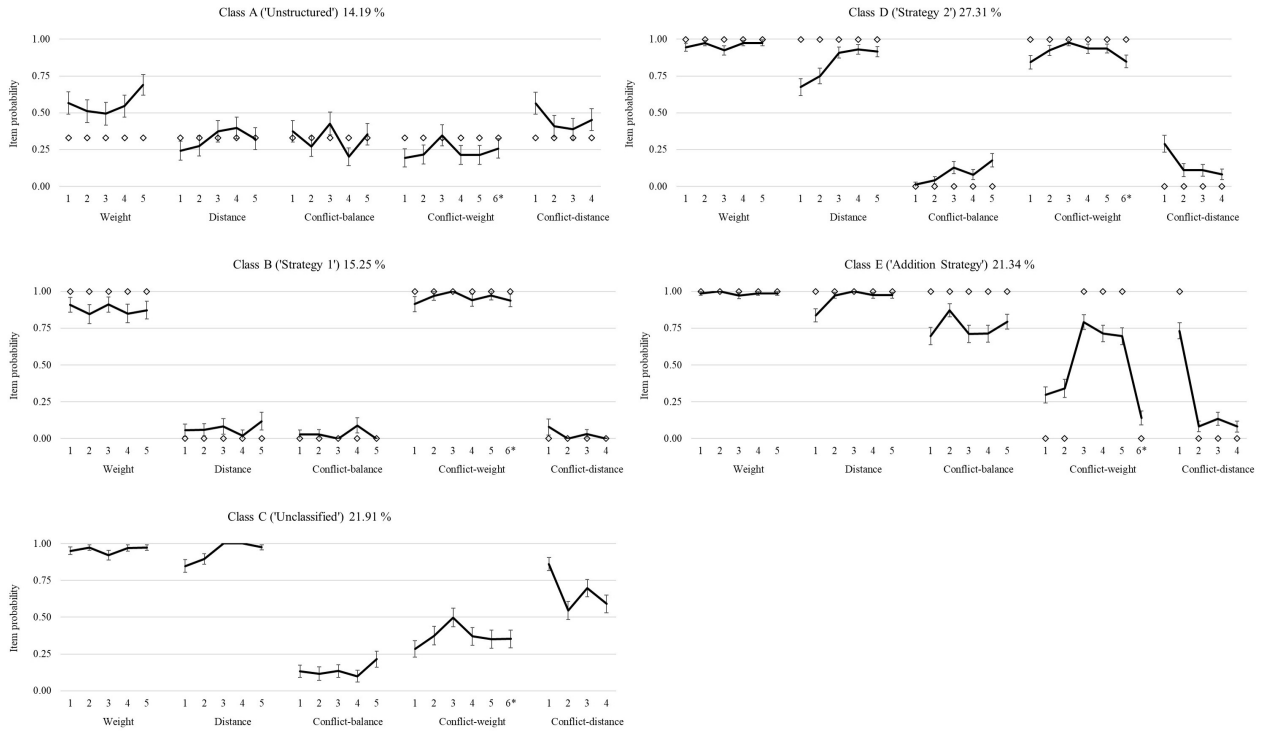


FIGURE 3 | Item probabilities for the six classes of the latent transition analysis (LTA), with the corresponding strategy and percentage of usage at pretest, and corresponding 90% confidence intervals. Items were distance, conflict-weight, conflict-distance, and conflict-balance; each had five items. *Note that item 5 of conflict-distance was digitized incorrectly and was a conflict-weight item.

children between 50% and 75%, and one below 50%. The one child with class assignment below 50% on posttest was assigned to class C (48%) and class E (48%), whose response patterns appeared to overlap.

The Transitions: Modeling Change

Change in strategy use was examined in more detail by investigating the transition from one class at pretest to another at posttest. Transition probabilities indicate the probability of moving from a specific class to another class; see **Table 5** for the transition probabilities for the control and BB conditions combined. Most class assignments did not change; see the diagonal. For example, 76% of class E (Addition Strategy) remained class E at posttest. Although there were some declines, mainly from Strategy 2 to Unclassified (20%), most changes in class indicated use of a better strategy, such as from Strategy 1 to Unclassified and from Unclassified to Addition Strategy. How these effects could be explained by condition was investigated next.

The number of children who transitioned was calculated for each condition. To do so, children were assigned to a single class, based on the largest probability to belong to a class. Similarly, transition probabilities per condition were requested and transformed into the number of children who transitioned. In both conditions, most children did not change their strategy; see **Figure 4**. In the BB condition, there were more transitions that indicated advancement in strategy compared to the control condition. In addition, the advancements indicated larger increases in accuracy on average, and the declines indicated smaller decreases in accuracy compared to the control condition. All transitions indicating improvement were present in the BB condition, except for the transition from Unclassified to Strategy 2. This cannot be seen in **Figure 4**, as small transitions were omitted for parsimony, but see **Appendix B** for all transitions per condition. To summarize, 65 children kept using the same strategy, 16 children declined, and 28 children improved in the BB condition.

TABLE 4 | Classes, their names, usage, and overall correct at pretest.

Class	Class name	Usage at pretest	Overall correct
A	Unstructured	14.19%	37.22%
B	Strategy 1	15.25%	42.87%
C	Unclassified	21.91%	60.51%
D	Strategy 2	27.31%	61.88%
E	Addition strategy	21.34%	69.92%

DISCUSSION

The aim of the present study was to examine conceptual change as a result of completing an inquiry-based lesson about balance beams compared to a control condition that entailed completing

TABLE 5 | Transition probabilities for classes from pretest to posttest.

		Posttest class				
		Unstructured	**Strategy 1**	**Unclassified**	**Strategy 2**	**Addition strategy**
Pretest class	Unstructured	0.49	0.12	0.22	0.11	0.06
	Strategy 1	0.09	0.25	0.28	0.25	0.14
	Unclassified	0.06	0	0.64	0	0.30
	Strategy 2	0.15	0.02	0.20	0.60	0.03
	Addition strategy	0.10	0	0.09	0.06	0.76

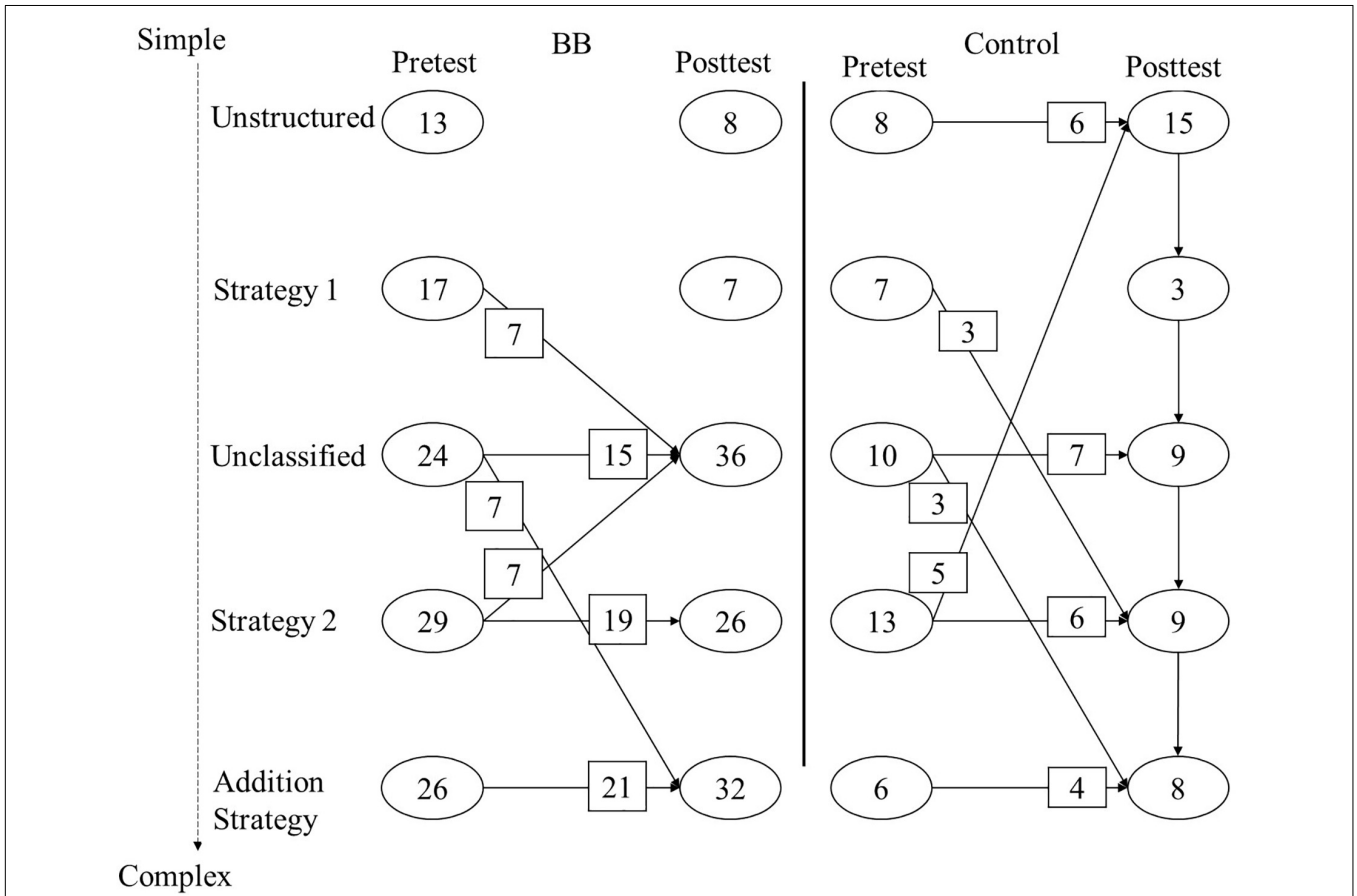


FIGURE 4 | Number of children per class for pretest and posttest with the transitions for the control and balance beam (BB) conditions. Transitions smaller than 5% of the sample were omitted from the figure for parsimony; see **Appendix B** for all transitions.

an inquiry-based lesson about gears. In particular, the changes in strategy use between pretest and posttest when solving balance beam problems were examined. The results showed that classes could be detected that generally matched the strategies proposed by Siegler (1976) and the classes found by Boom et al. (2001). In addition, children in the BB condition showed advances in understanding the law of moments, and no learning occurred in the control condition on average. Additional insights were obtained by detecting transitions between classes from pretest to posttest with LTA. LTA showed how condition affected learning: transitions to complex and accurate strategies occurred more often in the BB condition. Most improvements in the BB condition could be attributed to the adoption of the Addition Strategy (Wilkening and Anderson, 1982) and the "Unclassified" class, which showed some resemblance to Strategy 3. Together, these results showed that a single session of inquiry-based learning about the balance beam induced conceptual change in about 26% of the children, that conceptual change can be measured as advancement (or for some, a decline) in strategy use, and that this can be modeled using latent models.

The classes that were found corresponded reasonably well with the strategies proposed by Siegler (1976) and the classes found by

Boom et al. (2001). There were two classes that approximately matched the strategies proposed by Siegler (1976): classes B and D, which matched Strategies 1 and 2. Four out of six classes identified by Boom et al. (2001) were found in the present study. These were classes revealing Strategies 1 and 2 and the Addition Strategy, and an Unclassified class. Furthermore, these classes were used as often in the present study as by the 11.4-year olds in the study by Boom et al. (2001), with only the Unclassified class being used more often in the present study. Most children used Strategy 2 (27%), the Addition Strategy (21%), or the Unclassified strategy (22%). Taken together, the results show that Strategy 2 and the Addition Strategy are common strategies that can be identified in 9- to 11-year-old children.

Class A deviated most from previous studies. Class A ("Unstructured") showed a pattern that did not reveal consistency in the responses. The response pattern appeared to suggest guessing, which might be explained by randomly selecting an answer without carefully thinking. This could mean that children in this class did not know the answer on any of the items. Previous studies show at least understanding of weight as a factor in balancing beams (e.g., Boom et al., 2001). Class A did show accuracy above chance for weight. Therefore, children in this class did show some understanding, but the response pattern did not fit any strategy. An explanation of the low accuracy might be the young age of the present sample in combination with no previous experience with the law of moments.

The Unclassified class did not show a good match to Siegler (1976) strategies, but the present results support some overlap with Strategy 3: using both weight and distance but in an inconsistent manner. One result supporting the overlap with Strategy 3 is that items with weight only and distance only were answered correctly in this class, while there was large variation in the conflict items, where weight and distance have to be combined to answer the items correctly. Another reason is that there were no transitions to Strategy 2 (class D), where weight is the dominant variable in predicting which side goes down and distance is only used when weights are equal, but there were transitions to the Addition Strategy (class E). It can therefore be speculated that children using the Unclassified class had more advanced understanding than those using Strategy 2, even though their overall accuracy was slightly lower.

No class was found matching Strategy 4: perfect accuracy (Siegler, 1976). This might be explained by the age of the children in the present study: 10 years and 9 months on average. In another study, 11.4-year olds barely used Strategy 4: around 2% (Boom et al., 2001). It was expected that improved accuracy on the balance beam test as a result of the inquiry-based lesson about balance beams would also result in some children using Strategy 4. This did not happen, which suggested that acquisition of the most complex strategy in this age group is unlikely after a single inquiry-based lesson.

The present results showed that on average, children performed better on the balance beam test after the inquiry-based lesson about the balance beam than before, while those in the control condition did not improve. This means that a single session of inquiry-based learning could promote conceptual change, measured as a change in strategy use. This result can be explained by the effectiveness of inquiry-based learning (Lazonder and Harmsen, 2016). Others have also found beneficial effects of inquiry-learning (Huang et al., 2017) or problem-based learning (Loyens et al., 2015) on conceptual change. In the present study, the inquiry-based lesson consisted of elements that addressed activation of prior knowledge (hypothesis generation), critical analysis of findings (data interpretation), and deep comprehension (evidence evaluation). These elements were also found to be effective in problem-based learning (Loyens et al., 2015). The present inquiry-based lesson included additional supportive features: explication of predictions, self-generation of evidence, perceiving evidence, and verifying or falsifying understanding. These features target difficulties in conceptual change that were identified in primary school children (Vosniadou, 2002). The present study is one of the first to show that inquiry-based learning in primary school indeed fosters conceptual change.

A main result of this study was that improvement on the balance beam test could be attributed to specific transitions from one class to another, indicating that a different strategy was used after the inquiry-based lesson in the BB condition as compared to before. About 26% of the children in the BB condition advanced in strategy use. Most of these children who improved started with applying weight only (Strategy 1) and learned to use both weight and distance when weight was equal on both sides of the balance beam (Unclassified) or went from this Unclassified strategy to the Addition Strategy. While one improvement indicated an advancement of one step (Unclassified to Addition Strategy), there were numerous instances of improvements of more than one step, including from Strategy 1 to Unclassified, where Strategy 2 seems to be skipped. Note that although this is speculative, it showed that conceptual change does not have to be slow (Vosniadou, 2002). A different reason for change might be a different conceptualization of the problem and not a different conceptualization of the law of moments, which might be addressed in future studies. Nevertheless, this finding highlighted the advantage of using latent analyses, because overall improvement in the total score could be attributed to specific advancements in strategy use. In addition, the control condition did show transitions as well, although the total score did not change. It appeared that some children declined and some improved in strategy use. There might be multiple explanations for this finding. One explanation would be that children have access to multiple strategies at one time [see the overlapping waves model (Siegler, 2000)] and might switch to some extent from pretest to posttest. Another explanation is that responses to the problems are unstable because children do not have a coherent explanatory framework yet (Vosniadou, 2002).

Given that conceptual change occurred, it can be assumed that children experienced disequilibrium during inquiry-based learning. Disequilibrium is a moment in which the current explanatory framework is not adequate and children are open to new explanations (Piaget, 1978). After this moment, children can advance, or children can decline. The present results found both. In the case of a decline, new strategies entailed a less correct approach to the problems. This result resembled the results of Huang et al. (2017), who found that correct propositions

were adopted after inquiry-based learning but not assembled together correctly. In the present study, children did seem to correctly identify the need for a new strategy, but they did not yet create a coherent explanatory framework that could be used to solve the problems.

A final result regarding the transitions was that about 60% of the children in the BB condition did not move to another class and, thus, did not change strategy use. This is in line with the small-sized effect (Cohen, 1992) of the inquiry-based lesson. The effect, however, is close to the average found in educational interventions (Hattie, 2012). It is likely that the children who did not move to another class did not experience disequilibrium (Piaget, 1978) and/or were not able to notice other potential explanatory variables (Siegler and Chen, 1998). This might be due to, for example, children's scientific reasoning skills or the learning materials. It has been shown that the process of inquiry-based learning is performed better when children have strong scientific reasoning skills (Stender et al., 2018), but other factors that support scientific reasoning could affect the inquiry process as well, such as self-control or verbal skills (Van der Graaf et al., 2018), or Theory of Mind (Koerber and Osterhaus, 2019). Self-control and verbal skills might help to perform the most important process according Siegler and Chen (1998)—noticing potential explanatory variables—because it enables children to stop and think (self-control) and to access perceptual experiences via verbalization (verbal skills). On the other hand, the learning materials were effective overall but not for everyone. To improve learning for each child, the lesson might be repeated (Anderson, 1982), or additional support could be built in, for example, to reduce working memory load (Sweller et al., 2011), which could also lead to the acquisition of Strategy 4 in some children.

Exploratory analyses with age and gender were conducted. Gender did not relate to the pretest, posttest, or difference score (posttest minus pretest), which was in line with gender being not significant as a covariate in an LCA on the balance beam test (Boom and Ter Laak, 2007). The present results showed that gender also did not relate to a difference score. Age correlated positively with pretest scores and negatively with the difference score but only in the BB condition, in which children were slightly younger on average than in the control condition. Thus, in the younger group of children, young children appear to be less accurate in the balance beam test. This finding, again, is in line with previous studies on balance beam understanding (Boom et al., 2001; Boom and Ter Laak, 2007). What the present results add is that when learning occurs (i.e., in the BB condition), younger children tend to learn more. This result can best be interpreted with the relation between age and pretest scores in mind, which means that children with low pretest scores have large difference scores. Inquiry-based learning has previously been shown to result in larger learning gains for low-prior-knowledge children (e.g., Wang et al., 2010). As it is yet unknown what the effect of age on transitions from pretest to posttest would be after an inquiry-based lesson about the balance beam, it would be interesting to analyze latent classes, class assignment, and transitions, while using age as a covariate of class assignment, as Boom and Ter Laak (2007) did, and of transitions as well.

Some limitations apply to the present study. One limitation was the number of participants. As their response patterns were used to identify classes and this could change from pretest to posttest, more participants would be preferable to be better able to identify additional classes, especially classes that might have existed at pretest only or that emerged at posttest. There is no definite rule for sample size in LCA, because it also depends on circumstances, but sample size should be at least 70 participants (Hickendorff et al., 2018). The current number of participants is justified by classes consisting of at least 20 children, and classes were previously identified with larger sample sizes (Boom et al., 2001; Boom and Ter Laak, 2007). A better number of participants, especially when using covariates, would be around 500 participants (e.g., Boom and Ter Laak, 2007). Another limitation is that the actions of the learners were not recorded. It might be that children who designed the experiments as instructed learned more. A third limitation is that individual differences beyond age and gender were not taken into account. It is likely that the inquiry-based lesson required children to use their working memory (e.g., Kwon and Lawson, 2000) because inquiry-based learning activities involve, amongst others, use of prior knowledge and tracking manipulations of variables and experimental outcomes. In the present study, working memory demands were reduced by allowing hypotheses, results, and conclusions to be written down, allowing the revisiting of previously presented information, and more. Instruction in the digital learning environment was presented using text, which means that reading levels might have affected the results, possibly as a moderator (Graesser et al., 2005). However, reading levels were not assessed in the present study.

A strength of the present study was that transitions from pretest to posttest in strategy use when solving balance beam problems were uncovered. These transitions indicate that conceptual chance occurred. While more research is needed, the present results showed that it is feasible to assess conceptual change with the balance beam test after a short inquiry-based lesson about balance beams.

In the future, a feature might be added to prevent overloading working memory, namely, note-taking, especially when comparing the products (weight multiplied by distance) of the two sides of the balance beam. Note-taking (vs. not note-taking) might be especially advantageous for transfer to problem solving tests (Peper and Mayer, 1986). In addition, reviewing notes that were written down during the experiment positively relates to science understanding (Klein, 2000). An extended intervention could also be analyzed using LTA, which can show whether repeated exposure to the correct scientific theory induces additional advancement. LTA can reveal transitions between classes. LTA can also reveal which newly acquired strategy at posttest is most likely to be disregarded at a later time, if there is a decline in accuracy, as can happen in delayed posttests. It might be expected that a strategy corresponding to Strategy 3 or 4, where weight and distance are combined, is more difficult to learn and retain than Strategy 2, and therefore, it might be used at posttest but not anymore at delayed posttest. A final suggestion would be to investigate individual differences in strategy use and change. Covariates can be used in LTA to investigate this,

for example, understanding of a correct experimental strategy positively related to knowledge about floating and sinking and to the probability to change to a more proficient knowledge profile (Edelsbrunner et al., 2018).

The results suggested that for classroom science education, it is important to take into account differences between and within individuals at a specific time and over time (Hickendorff et al., 2018). The classes showed the existence of multiple strategies in a group of children (Jansen and van der Maas, 1997). Some classes suggested a mixture of strategies being present at one time (Siegler, 2000). Some children improved their strategy use when predicting which side of a balance beam goes down. Therefore, a suggestion would be to present digital inquiry-based learning activities in which the learning process is recorded to get insight into explanatory frameworks in use (cf. diSessa, 2002), to get insight into optimal learning moments (disequilibrium; Piaget, 1978), and to provide teachers with potential learning paths for the heterogeneous group of children being taught. Another suggestion is to focus on how to combine weight and distance to correctly calculate momentum. In the present study, most children in the BB condition understood that both variables matter at posttest but did not know how to combine them correctly.

To conclude, strategy use when solving problems with the balance beam could be identified and the classification reasonably matched with the strategies proposed by Siegler (1976). The strategy that was used changed after an inquiry-based lesson about balance beams for about 40% of the children, indicating that conceptual change occurred. It is, therefore, feasible to assess conceptual change with the balance beam test after inquiry-based learning in primary school.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by BMS Ethics Committee, University of Twente. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

I would like to thank Evelien Mulder, Inge Molenaar, and Ard Lazonder for their useful feedback on an earlier version of this manuscript and Peter Edelsbrunner for his useful comments on the analyses.

REFERENCES

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychol. Rev.* 89, 369–406. doi: 10.1037/0033-295X.89.4.369

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2014). Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. *Test* 23, 433–465. doi: 10.1007/s11749-014-0381-7

Bartolucci, F., Pandolfi, S., and Pennoni, F. (2017). LMest: an R package for latent markov models for longitudinal categorical data. *J. Stat. Softw.* 81, 1–38. doi: 10.18637/jss.v081.i04

Boom, J., Hoijtink, H., and Kunnen, S. (2001). Strategies in the balance classes, strategies, or strategies for the balance scale test? *Cogn. Dev.* 16, 717–735. doi: 10.1016/S0885-2014(01)00056-9

Boom, J., and Ter Laak, J. (2007). Classes in the balance: latent class analysis and the balance scale test. *Dev. Rev.* 27, 127–148. doi: 10.1016/j.dr.2006.06.001

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Celeux, G., and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *J. Classification* 13, 195–212. doi: 10.1007/bf01246098

Chi, M. T. H., and Roscoe, R. D. (2002). "The processes and challenges of conceptual change," in *Reconsidering Conceptual Change*, eds M. Limón and L. Mason (Dordrecht: Kluwer Academic Publishers).

Chinn, C. A., and Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *J. Res. Sci. Teach.* 35, 623–654. doi: 10.1002/(sici)1098-2736(199808)35:6<623::aid-tea3>3.0.co;2-o

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155–159. doi: 10.1037/0033-2909.112.1.155

diSessa, A. A. (2002). "Why "conceptual ecology" is a good idea," in *Reconsidering Conceptual Change*, eds M. Limón and L. Mason (Dordrecht: Kluwer Academic Publishers).

diSessa, A. A., and Sherin, B. L. (1998). What changes in conceptual change? *Int. J. Sci. Educ.* 20, 1155–1191. doi: 10.1080/0950069980201002

Dunbar, K. N., and Klahr, D. (2012*). "Scientific thinking and reasoning," in *The Oxford Handbook of Thinking and Reasoning*, eds K. J. Holyoak and R. G. Morrison (Oxford: Oxford University). doi: 10.1093/oxfordhb/9780199734689.013.0035

Edelsbrunner, P. A., Schalk, L., Schumacher, R., and Stern, E. (2018). Variable control and conceptual change: a large-scale quantitative study in elementary school. *Learn. Individ. Differ.* 66, 38–53. doi: 10.1016/j.lindif.2018.02.003

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., et al. (2014). Scientific reasoning and argumentation: advancing an interdisciplinary research agenda in education. *Front. Learn. Res.* 5:28–45. doi: 10.14786/flr.v2i3.96

Graesser, A. C., McNamara, D. S., and VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query. AutoTutor, and iSTART. *Educ. Psychol.* 40, 225–234. doi: 10.1207/s15326985ep4004-4

Hattie, J. (2012). *Visible Learning for Teachers: Maximizing Impact on Learning*. New York, NY: Routledge.

Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., and Trezise, K. (2018). Informative tools for characterizing individual differences in learning: latent class, latent profile, and latent transition analysis. *Learn. Individ. Differ.* 66, 4–15. doi: 10.1016/j.lindif.2017.11.001

Huang, K., Ge, X., and Eseryel, D. (2017). Metaconceptually-enhanced simulation-based inquiry: effects on eight grade students' conceptual chance and epistemic beliefs. *Educ. Technol. Res. Dev.* 65, 75–100. doi: 10.1007/s11423-016-9462-5

Inhelder, B., and Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence: an Essay on the Construction of Formal Operational Structures*. New York, NY: Basic Books. (A. Parsons & S. Milgram, Trans.)

Jansen, B. R. J., and van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Dev. Rev.* 17, 321–357. doi: 10.1006/drev.1997.0437

Jansen, B. R. J., and van der Maas, H. L. J. (2002). The development of children's strategy use on the balance scale test. *J. Exp. Child Psychol.* 81, 383–416. doi: 10.1006/jecp.2002.2664

Klahr, D. (2000). *Exploring Science: the Cognition and Development of Discovery Processes*. Cambridge, MA: MIT Press.

Klein, P. D. (2000). Elementary students' strategies for writing-to-learn in science. *Cogn. Instruct.* 18, 317–348. doi: 10.1207/S1532690XCI1803-2

Kline, P. (1993). *The Handbook of Psychological Testing*. London: Routledge.

Koerber, S., and Osterhaus, C. (2019). Individual differences in early scientific thinking: assessment, cognitive influences, and their relevance for science learning. *J. Cogn. Dev.* 20, 510–533. doi: 10.1080/15248372.2019.1620232

Koerber, S., Osterhaus, C., and Sodian, B. (2017). Diagrams support revision of prior belief in primary-school children. *Front. Learn. Res.* 5:76–84. doi: 10.14786/flr.v5i1.265

Kwon, Y., and Lawson, A. E. (2000). Linking brain growth with the development of scientific reasoning ability and conceptual change during adolescence. *J. Res. Sci. Teach.* 37, 44–62. doi: 10.1002/(sici)1098-2736(200001)37:1<44::aid-tea4>3.0.co;2-j

Lazonder, A. W., and Harmsen, R. (2016). Meta-analysis of inquiry-based learning: effects of guidance. *Rev. Educ. Res.* 86, 681–718. doi: 10.3102/0034654315627366

Loyens, S. M. M., Jones, S. H., Mikkers, J., and Van Gog, T. (2015). Problem-based learning as a facilitator of conceptual change. *Learn. Instruct.* 38, 34–42. doi: 10.1016/j.learninstruc.2015.03.002

Mussweiler, T., and Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: a selective accessibility model. *J. Exp. Soc. Psychol.* 35, 136–164. doi: 10.1006/jesp.1998.1364

Nyland, K. L., Asparouhov, T., and Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct. Equat. Model. Multidiscip. J.* 14, 535–569. doi: 10.1080/10705510701575396

Peper, R. J., and Mayer, R. E. (1986). Generative effects of note-taking during science lectures. *J. Educ. Psychol.* 78, 34–38. doi: 10.1037/0022-0663.78.1.34

Piaget, J. (1978). *The Development of Thought: Equilibration of Cognitive Structures*. Oxford, UK: Blackwell. (A. Rosin, Trans.).

Posner, G. J., Strike, K. A., Hewson, P. W., and Gertzog, W. A. (1982). Accommodation of a scientific conception: toward a theory of conceptual change. *Sci. Educ.* 66, 211–227. doi: 10.1002/sce.3730660207

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R foundation for statistical computing.

Schneider, M., and Hardy, I. (2013). Profiles of inconsistent knowledge in children's pathways of conceptual change. *Dev. Psychol.* 49, 1639–1649. doi: 10.1037/a0030976

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognit. Psychol.* 8, 481–520. doi: 10.1016/0010-0285(76)90016-5

Siegler, R. S. (2000). The rebirth of children's learning. *Child Dev.* 71, 26–35. doi: 10.1111/1467-8624.00115

Siegler, R. S., and Chen, Z. (1998). Developmental differences in rule learning: a microgenetic analysis. *Cogn. Psychol.* 36, 273–310. doi: 10.1006/cogp.1998.0686

Stender, A., Schwichow, M., Zimmerman, C., and Härtig, H. (2018). Making inquiry-based science learning visible: the influence of CVS and cognitive skills on content knowledge learning in guided inquiry. *Int. J. Sci. Educ.* 40, 1812–1831. doi: 10.1080/09500693.2018.1504346

Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive Load Theory*. New York, NY: Springer.

Van der Graaf, J., Segers, E., and Verhoeven, L. (2018). Individual differences in the development of scientific thinking in kindergarten. *Learn. Instruct.* 56, 1–9. doi: 10.1016/j.learninstruc.2018.03.005

Vosniadou, S. (2002). "On the nature of naïve physics," in *Reconsidering Conceptual Change*, eds M. Limón and L. Mason (Dordrecht: Kluwer Academic Publishers).

Wang, J., Wang, Y., Tai, H., and Chen, W. (2010). Investigating the effectiveness of inquiry-based instruction on students with different prior knowledge and reading abilities. *Int. J. Sci. Math. Educ.* 8, 801–820. doi: 10.1007/s10763-009-9186-7

Wilkening, F., and Anderson, N. H. (1982). Comparison of two strategy-assessment methodologies for studying cognitive development and knowledge structure. *Psychol. Bull.* 92, 215–237. doi: 10.1037/0033-2909.92.1.215

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 van der Graaf. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX A

TABLE A1 | Item characteristics: number, type, weight, and distance on the left and right sides of the balance beam, torque, and which side goes down.

		Item configuration						
		Left		Right				
Item number	**Item type**	***D***	***W***	***D***	***W***	**Torque**	**Side down**	**Addition strategy**
1	W	2	1	2	2	2	Right	Correct
2	W	3	3	3	4	3	Right	Correct
3	W	1	2	1	4	2	Right	Correct
4	W	2	3	2	2	−2	Left	Correct
5	W	4	3	4	1	−8	Left	Correct
6	D	2	1	1	1	−1	Left	Correct
7	D	3	2	2	2	−2	Left	Correct
8	D	4	3	2	3	−6	Left	Correct
9	D	3	4	4	4	4	Right	Correct
10	D	2	3	3	3	3	Right	Correct
11	CB	1	4	4	1	0	Balance	Correct
12	CB	3	1	1	3	0	Balance	Correct
13	CB	3	2	2	3	0	Balance	Correct
14	CB	3	4	4	3	0	Balance	Correct
15	CB	4	2	2	4	0	Balance	Correct
16	CW	4	2	3	3	1	Right	Incorrect
17	CW	2	3	4	1	−2	Left	Incorrect
18	CW	3	1	2	3	3	Right	Correct
19	CW	1	4	3	1	−1	Left	Correct
20	CW	3	2	2	4	2	Right	Correct
21	CD	1	4	3	3	5	Right	Correct
22	CD	2	2	1	3	−1	Left	Incorrect
23	CD	3	3	2	4	−1	Left	Incorrect
24	CD	1	4	2	3	2	Right	Incorrect
25	CW*	4	2	3	3	1	Right	Incorrect

*W, weight; D, distance; C, conflict. *This CW item should have been a CD item but was digitized incorrectly. The number of weights is presented, and each weight was 5 kg.*

APPENDIX B

TABLE B1 | Number of children in the control condition who transitioned from pretest to posttest.

		Posttest class				
		Unstructured	**Strategy 1**	**Unclassified**	**Strategy 2**	**Addition strategy**
Pretest class	Unstructured	6	2	0	0	0
	Strategy 1	2	1	0	3	1
	Unclassified	0	0	7	0	3
	Strategy 2	5	0	2	6	0
	Addition strategy	2	0	0	0	4

TABLE B2 | Number of children in the balance beam (BB) condition that transitioned from pretest to posttest.

		Posttest class				
		Unstructured	**Strategy 1**	**Unclassified**	**Strategy 2**	**Addition strategy**
Pretest class	Unstructured	5	1	4	2	1
	Strategy 1	0	5	7	3	2
	Unclassified	2	0	15	0	7
	Strategy 2	1	1	7	19	1
	Addition strategy	0	0	3	2	21

frontiers in Psychology

ORIGINAL RESEARCH
published: 19 November 2020
doi: 10.3389/fpsyg.2020.531503

Preschoolers' Induction of the Concept of Material Kind to Make Predictions: The Effects of Comparison and Linguistic Labels

Ilonca Hardy[1], Henrik Saalbach[2,3], Miriam Leuchter[4] and Lennart Schalk[5]*

[1] Institute of Early and Primary Education, Educational Sciences, Goethe University Frankfurt, Frankfurt, Germany, [2] Institute of Educational Sciences, Leipzig University, Leipzig, Germany, [3] Leipzig Research Center for Early Childhood Development, Leipzig University, Leipzig, Germany, [4] Institute for Children and Youth Education, Educational Sciences, University of Koblenz and Landau, Landau, Germany, [5] Institute for Research on Instruction and Subject-Specific Didactics, PH Schwyz, Goldau, Switzerland

Analogical reasoning by comparison is considered a special case of inductive reasoning, which is fundamental to the scientific method. By reasoning analogically, learners can abstract the underlying commonalities of several entities, thereby ignoring single objects' superficial features. We tested whether different task environments designed to trigger analogical reasoning by comparison would support preschoolers' induction of the concept of material kind to predict and explain objects' floating or sinking as a central aspect of scientific reasoning. Specifically, in two experiments, we investigated whether the number of presented objects (one versus two standards), consisting of a specific material and the labeling of objects with the respective material name, would benefit preschoolers' material-based inferences. For each item set used in both experiments, we asked the children (N = 59 in Experiment 1, N = 99 in Experiment 2) to predict an object's floating or sinking by matching it to the standards and to verbally explain their selections. As expected, we found a significant effect for the number of standards in both experiments on the prediction task, suggesting that children successfully induced the relevance of material kind by comparison. However, labels did not increase the effect of the standards. In Experiment 2, we found that the children could transfer their conceptual knowledge on material kind but that transfer performance did not differ among the task environments. Our findings suggest that tasks inviting analogical reasoning by comparison with two standards are useful for promoting young children's scientific reasoning.

Keywords: scientific reasoning, comparison, induction, preschool, labeling

OPEN ACCESS

Edited by:
Ageliki Nicolopoulou, Lehigh University, United States

Reviewed by:
Katharina J. Rohlfing, University of Paderborn, Germany
Catherine Sandhofer, University of California, Los Angeles, United States

****Correspondence:***
Ilonca Hardy
hardy@em.uni-frankfurt.de

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *31 January 2020*
Accepted: *28 September 2020*
Published: *19 November 2020*

Citation:
Hardy I, Saalbach H, Leuchter M and Schalk L (2020) Preschoolers' Induction of the Concept of Material Kind to Make Predictions: The Effects of Comparison and Linguistic Labels. Front. Psychol. 11:531503. doi: 10.3389/fpsyg.2020.531503

INTRODUCTION

Analogical reasoning by comparison is assumed to be a crucial mechanism, enabling induction and conceptual learning across different age groups and in a wide range of tasks (Loewenstein and Gentner, 2001; Alfieri et al., 2013; Schalk et al., 2016). According to the theory of structural alignment (Gentner, 2010), analogical reasoning involves individuals' identification, mapping,

and evaluation of the similarities and differences of several entities. This process has benefits for building conceptual knowledge because it supports individuals' encoding of information, induction and abstraction of categories, and generalization (transfer) of knowledge (Gentner and Smith, 2012; Gentner and Hoyos, 2017). While analogical reasoning by comparison has been investigated in experimental settings on a range of conceptual learning tasks (Alfieri et al., 2013), its contribution to scientific reasoning is rarely considered. We propose that task environments that trigger comparison are relevant to scientific reasoning. Specifically, such environments may facilitate children's induction of scientific concepts as a basis for predictions and explanations by their encoding of relevant object features.

In the present research, we investigated different task environments that may facilitate preschoolers' encoding of relevant object features to generate predictions and explanations in the science context of "floating and sinking." Research with preschoolers across various science contexts has revealed that preschoolers typically hold a variety of naïve conceptions based on irrelevant and perceptually salient features and that these conceptions may affect hypothesis generation (Carey, 2000). For example, when preschoolers are asked to predict the floating or sinking of solid objects and to explain their predictions, they provide explanations such as "light things will float," "things with holes will sink," "large things will sink," or "things with air in them will float" (Penner and Klahr, 1996; Leuchter et al., 2014). Thus, preschoolers' predictions about whether objects float or sink are typically based on salient features, such as weight, size, and form, rather than on generic and more abstract aspects such as material kind or density. Children's naïve conceptions prevail despite the fact that they are typically able to name the material of solid objects, such as wood, plastic, or metal (Smith et al., 1985; Dickinson, 1987; Leuchter et al., 2014). Typically, a process of conceptual restructuring is required for children to overcome their naïve conceptions and transform them into scientifically advanced conceptions (Schneider et al., 2012; Leuchter et al., 2014). In two experiments, we investigated whether different task environments intended to trigger analogical reasoning by comparison (i.e., presenting two objects at the same time and labeling them) would improve preschoolers' predictions and explanations of floating or sinking based on the concept of material kind.

Scientific Reasoning in Preschoolers

The goals of science education encompass mastery of scientific concepts as well as learning how to engage in scientific reasoning (Driver et al., 2009; Kuhn, 2010; Dunbar and Klahr, 2012; Sandoval et al., 2014). In general terms, scientific reasoning involves individuals' knowledge-seeking by the application of scientific methods (Klahr and Dunbar, 1988; Klahr, 2000; Piekny and Maehler, 2013). Models of scientific reasoning typically refer to processes of inductive reasoning to explain individuals' knowledge construction, as well as their hypothesis and inference generation (Zimmerman, 2007; Morris et al., 2012). Inductive reasoning is regarded as a cognitive process that captures how individuals encode information, mentally represent this information, organize information into patterns, and derive inferences (cf. Chinn and Brewer, 2001; Chinn and Malhotra, 2002; Dunbar and Klahr, 2012). With respect to hypothesis generation, Klahr and Dunbar (1988) differentiate between "evoking" and "inducing" a hypothesis. When evoking a hypothesis, individuals retrieve and rely on prior knowledge. When inducing a hypothesis, individuals need to observe and encode data, and to identify patterns, before venturing an initial hypothesis. Therefore, the generation of a hypothesis requires children to encode relevant observations, to identify underlying patterns, and to draw inferences, which may rely in part on prior conceptual knowledge.

Several reviews indicate that even preschoolers can exhibit basic aspects of scientific reasoning (Chinn and Malhotra, 2002; Zimmerman, 2007; van der Graaf et al., 2016). There is substantial evidence indicating that preschoolers can appropriately generate hypotheses, identify common patterns in data, and evaluate presented evidence in specific task contexts. With regard to hypothesis generation, Piekny and Maehler (2013) found that 4- to 6-year-olds were able to construct a hypothesis based on patterns of evidence. If task contexts are kept simple, 5- and 6-year-olds are even able to identify patterns that are more complex and to form various hypotheses (e.g., Sodian et al., 1991; Ruffman et al., 1993). Moreover, an intervention study by Schulz et al. (2007) revealed that 4- to 6-year-olds were not only able to infer the causal structure of events by using experimentally collected patterns of evidence but also to predict the outcome of a novel experiment. Piekny et al. (2014) found that 4- to 6-year-olds were able to evaluate conclusive and partially conclusive evidence correctly, and Koerber et al. (2005) found that the correspondence between preschoolers' conceptions and presented evidence facilitated evaluation, whereas conflicting conceptions impeded evaluation. Even the 4-year-olds in the study were able to understand that data with perfect covariation could corroborate or disconfirm a causal hypothesis (see also Tullos and Woolley, 2009). Finally, a study by Köksal-Tuncer and Sodian (2018) revealed that 3- to 6-year-olds were not only able to generate hypotheses but also to apply hypothesis-testing strategies when presented with counterevidence.

In this prior research, selection tasks and production tasks were employed to assess young children's generation of hypotheses in scientific reasoning contexts (Koerber et al., 2005; Piekny and Maehler, 2013; Gropen et al., 2017). Selection tasks provide children with different answer options to a given problem and may be employed to assess children's spontaneous reasoning with regard to a given scientific reasoning context. By contrast, production tasks require children to come up with solutions themselves on the basis of explicit reasoning. Production tasks have therefore been employed to assess children's deliberate reasoning based on the production, explanation, and evaluation of arguments (Mercier and Sperber, 2009; Mercier, 2011). For our two experiments, we used a selection task in which we asked children to predict which object would float or sink like one or two other objects (prediction task). We also employed a production task in which we asked children to explain their respective prediction in order to assess their deliberate reasoning (explanation task).

Overall, preschoolers' success in generating hypotheses is typically assessed in terms of the adequacy of their predictions and explanations based on given data patterns. However, research has rarely focused on the specific task conditions supporting the formation of predictions and explanations. We suggest that research on analogical reasoning by comparison (Gentner and Smith, 2012; Alfieri et al., 2013) may provide insights and that task contexts triggering analogical reasoning may benefit the performance of young children in a scientific reasoning context.

Analogical Reasoning by Comparison

Analogical reasoning can be understood as a special case of inductive reasoning (Holland et al., 1989; Holyoak, 2005) because it refers to individuals' ability to integrally encode commonalities and differences across a variety of entities and situations, represent and re-represent this information, and draw inferences (Gentner, 2016). Analogical reasoning has been demonstrated to be fundamental to cognitive development and conceptual learning, such as spatial orientation, word learning, learning of principles, and social comparison (for summaries, see Holyoak, 2005; Gentner, 2010; Gentner and Hoyos, 2017). One strategy for promoting analogical reasoning is to invite comparisons. According to the theory of structural alignment, comparison involves the retrieval of relevant information from the long-term memory, the mapping of commonalities and differences of two (or more) presented entities or cases (thereby inducing an abstracted schema), and the projection of inferences based on this mapping (Gentner and Smith, 2012). The abstracted schema will be more general than the analogs because inferences are formed on the basis of aligned similarities and differences, leading to a merged representation—a process of "relational pattern completion" (Gentner and Smith, 2012).

Schema abstraction can be supported by simultaneously presenting two entities or examples (Gentner and Medina, 1998; Gentner and Namy, 1999; Gentner et al., 2007; Kurtz and Loewenstein, 2007; Gentner, 2010; Alfieri et al., 2013; Christie, 2020). Specifically, Gentner and Namy (1999) argue that the presentation of two standards may "promote the discovery of relatively abstract relational commonalities that could characterize the category being learned" (p. 506). Christie (2020) provides empirical support to this claim. She reports on a series of studies in which young children successfully learned categories when presented with multiple exemplars. In addition, Christie and Gentner (2010) showed that 4-year-olds were able to recognize similarity in relations (above, under) in pictures of animals in different positions by making correct choices in a selection task. Gentner and Rattermann (1991) showed, in a more complex task, that 3-year-olds were unable to carry out relational matches without additional support, evidenced by choosing an object match instead of a relational match, but that 5-year-olds were successful in this task. Despite positive evidence from prompting comparisons even in young children, presenting two cases, examples, entities, or situations simultaneously may not always be sufficient for learners to recognize similarities (Kurtz et al., 2001). In particular, it has been shown that young children benefit from additional prompts, such as the use of common labels, when presented with two objects to be compared (Gentner, 2010; Christie, 2020).

Using Language to Promote Comparison

Labels and other types of verbal scaffold often facilitate analogical reasoning by comparison (Alfieri et al., 2013). Gentner and Namy (1999, 2006), Namy and Gentner (2002), Gentner (2010, 2016), and Hespos et al. (2020) argue that language, in general, plays a decisive role in triggering analogical reasoning by comparison even in young children. Language and structure-mapping are suggested to bootstrap each other, mutually influencing cognitive and conceptual development in young and older children as well as adults (Gentner, 2010; Christie, 2020). As Gentner and Rattermann (1991, p. 260) put it, "a word can function as a promissory note, signaling subtle commonalities that the child does not yet perceive." Gentner (2010) proposes four ways in which language and structure-mapping interact: (1) Common labels invite comparison and abstraction by highlighting similarity across entities; (2) the naming of entities promotes reification since it preserves abstraction linguistically; (3) the naming of entities promotes uniform relational encoding; and (4) the use of linguistic structures invites the construction of conceptual structures. In the present research, we focus on the function of labels to invite and trigger comparison.

In a series of experiments, Gentner et al. (2007) investigated the use of common labels for triggering analogical reasoning by comparison. For example, Namy and Gentner (2002) tested two groups of 4-year-olds using a forced-choice match-to-sample task in which the children had to extend the label of one object (the so-called *standard*) to one of two other objects. In the no-comparison group, the experimenter labeled a single standard (e.g., a picture of an apple) with a made-up name (e.g., blicket). The children had to decide which one of the two other objects, the taxonomic item (banana) or the perceptually similar item (balloon), would have the same name as the standard. In the comparison group, the task was the same except that the experimenter showed and labeled pictures of two taxonomically related standards (e.g., an apple and a pear) with the same made-up name. Namy and Gentner found that preschoolers' correct taxonomic choices for this task increased when two standards were presented rather than just one standard (see also Gentner and Namy, 1999). Importantly, children were even more likely to make taxonomic choices when the two standards were labeled with the same noun. By contrast, when the two standards received different labels, children did not engage in comparison, as indicated by their increased selection of the perceptually similar but taxonomically unrelated item.

This effect of labeling not only holds for children's category learning but also for their generalization of properties across entities. That is, the presence of a common label enhances young children's willingness to make inductive inferences between entities (e.g., Gelman and Markman, 1986; Gelman et al., 1986; Davidson and Gelman, 1990; Saalbach et al., 2012). For example, Gelman and Markman (1986) found that preschoolers can generalize properties across members of the same category when category membership is labeled with the same noun, but not when it is unlabeled. Thus, linguistic labels can serve as simple

scaffolds to trigger young children's comparison processes since they suggest similarity between two items.

Importantly, these beneficial effects of labels have typically been found in tasks that require only a low degree of prior knowledge in young children (Christie, 2020). In our study, we investigated the role of labels in a domain in which learners start with some degree of prior conceptual knowledge. Specifically, we propose that labels of material kind (e.g., "this is made of wood") will function as a cue at a superordinate level. Material labels provide rich associations with unique properties of objects, such as their texture and specific weight, as well as their empirically observable behavior of floating or sinking in water. From early on, children learn that common labels are used for things that are alike. If labels are used in an instructional context, they can elicit comparison of the respective objects and thereby highlight the relevant (underlying) commonality of different yet related entities even in the absence of perceptual similarity (Gentner, 2010). In several studies it has been shown that nouns provide information about classes of objects better than verbs or adjectives do (Arunachalam and Waxman, 2010; Graham et al., 2012). In addition, Johanson and Papafragou (2016) found that labeling using nouns works with similar success as labeling using facts (e.g., descriptions of properties) in ambiguous situations. In our two experiments, we systematically varied the use of superordinate-level nouns when labeling objects to promote comparison.

The Present Research

We tested whether different task environments designed to trigger analogical reasoning by comparison with or without labeling the respective objects would support preschoolers' induction of the concept of material kind to predict and explain objects' floating or sinking. Specifically, 4- to 7-year-old preschoolers were exposed to sets of material within a forced-choice match-to-sample task with variations in the number of standard objects and the use of superordinate labels. On the basis of Gentner (2010) and the results from Namy and Gentner (2002), we expected that triggering comparison via the use of linguistic labels would amplify the effect of presenting two objects simultaneously. In order to probe the effect of labeling on performance, we varied the extent to which the labels were employed across our two experiments.

EXPERIMENT 1

In Experiment 1, we crossed two factors: the number of standard objects and the use of material labels. Specifically, we presented either one standard or two standards to test the effect of comparison on children's predictions and explanations, and we either labeled the presented objects or did not label them at all in order to test the additional effect of labeling on the potential benefits of comparison. The labels referred to material kind and therefore served as superordinate category labels indicating common properties of the presented objects. Consequently, it was possible that they might amplify children's perception of the similarities between the presented objects in conditions with two standards. Before the children were randomly assigned to the four conditions of Experiment 1, their prior knowledge was assessed with a pretest and a baseline assessment.

Our research questions and hypotheses were as follows:

(1) Will preschoolers induce the concept of material kind when analogical reasoning by comparison is triggered by presenting two standards with the same floating behavior?
Hypothesis (1): Preschoolers in conditions with two standards will outperform preschoolers in conditions with one standard on the Prediction and Explanation Tasks.

(2) Does the use of material labels facilitate preschoolers' induction of the concept of material kind?
Hypothesis (2): The use of material labels will improve performance in the Prediction and Explanation Tasks in the condition with two standards, as indicated by an interaction effect of Labeling and Number of Standards.

Method

Participants

Fifty-nine preschoolers from a major German city who had German as a first language and a mean age of 5 years, 3 months (min. = 4 years, 11 months; max. = 6 years) participated in this study (testing was in German). They were recruited through preschools. Parental consent on participation was collected for all children. The children came from middle-class families living in urban and suburban areas.

Design

In a 2 × 2 between-groups design, we tested the importance of triggering a comparison (one standard or two standards) and the benefit of labeling standards with the respective material label (unlabeled versus labeled). Specifically, the four conditions were (1) One Unlabeled Standard (one_unlabeled); (2) One Labeled Standard (one_labeled); (3) Two Unlabeled Standards (two_unlabeled); (4) Two Labeled Standards (two_labeled). In the unlabeled conditions, the standards were referred to as *"this"/"these"*; in the labeled conditions, the standards' material was named (e.g., *"this/these is/are made of wood"*). In all conditions, children had to predict (Prediction Task) and explain (Explanation Task) which of the four selection objects would float or sink like the standards. **Table 1** gives an overview of the experimental conditions and the respective instructions.

Before the children were randomized to these four conditions, all children participated in a pretest and a baseline assessment. In the pretest, children had to match objects made of the same material (Matching Task) and to name the objects' material (Labeling Task). Subsequently, in the baseline assessment, children were presented with one standard that was not labeled and four selection objects. As in the four conditions described above, children had to predict (Prediction Task) and explain (Explanation Task) which of the four selection objects would float or sink like the standard.

Tasks and Procedure

All participants were tested individually in a quiet room in their preschools by an experimenter blind to the hypotheses.

TABLE 1 | Conditions and instructions in Experiment 1.

	One Standard	Two Standards
Unlabeled	Look, this one floats/sinks in water. Which of these also floats/sinks just like this one?	Look, these two float/sink in water. Which of these also floats/sinks just like these ones?
Labeled (example: wood)	Look, this is made out of wood and it floats in water. Which of these also floats just like this one: this one made of wood, this one made of metal, this one made of metal, or this one made of glass?	Look, this one is made out of wood and it floats in water. And this one is also made out of wood and it floats, too. Which of these also floats just like these ones: this one made of wood, this one made of clay, this one made of clay, or this one made of metal?

Testing began with the pretest and the baseline assessment, comprising the Prediction and Explanation Tasks; afterward, the children took the Prediction and Explanation Tasks with different materials, depending on the conditions.

Pretest: Matching/labeling task

Two tasks were employed to assess the children's prior knowledge of materials and their labels. The children were presented with a total of 16 objects made of eight different materials (wood, stone, metal, plastic, Styrofoam, wax, glass, clay), each with two different shapes (e.g., a wooden block and a wooden spoon). In the Matching Task, the children were asked to match pairs of objects: *"Find the two things that belong together and put them together on the table."* After a child had matched all the objects, the experimenter rearranged them in pairs by material regardless of how the children had arranged them in the Matching Task. In the Labeling Task, the children were asked to name the material: *"Tell us what the objects are made of."* The experimenter did not use the term "material." The Matching Task was scored with respect to the successful matching of objects according to their material, with one point assigned for each correct material-based match of two objects (range of scores 0–8). The Labeling Task was scored with one point for each correct material label (range of scores 0–8).

Baseline assessment: Prediction and explanation task

The baseline assessment served to measure how children would predict and explain objects' floating or sinking if they were only provided with a single standard that was not labeled, i.e., a task environment without elements to support comparison. Specifically, we employed six object sets in the baseline assessment. All objects in these sets were different from the objects used in the pretest. In each set, one object of a specific material served as the standard; four objects served as the selection items. In each object set, only one of the selection objects was made of the same material as the standard, but it always had a different shape and size. The other three selection objects were distractors that were selected on the basis of children's typical misconceptions (Hardy et al., 2006). Of the three distractors, there was one with the same shape as the standard. The other two distractors had a salient size or weight. That is, if the standards floated, we used extremely light and/or small selection objects (e.g., a small needle), whereas if the standards sank, we used extremely heavy and/or large selection objects (e.g., a large piece of wax). Only the selection object made of the same material as the standards sank/floated like the standards. Half of the sets had a standard made of material that floats in water (wood, wax, Styrofoam), and the other half had a standard made of material that sinks in water (metal, plastic, clay). **Table 2** provides an overview of all objects sets used in the baseline assessment and the conditions of Experiment 1 and Experiment 2.

In the Prediction Task the experimenter first took the standard and said "*Look, this floats/sinks in water.*" The experimenter then immersed the standard in a water basin, and the child observed whether the object sank or rose to the top. The experimenter then positioned this object above the four selection objects and asked "*Which of these also floats/sinks, just like this one here? This one, this one, this one, or this one?*" while pointing at the four selection items. The child then selected one of four objects that would float/sink just like the standard. The children received one point for choosing the selection object of the same material as the presented standard and zero points for choosing one of the other three selection objects (i.e., the range of possible scores for the Prediction Task was 0–6).

The Explanation Task followed immediately after the child had made their prediction. That is, once the child had chosen one of the four selection objects, the experimenter asked the child to explain their prediction by asking, *"Why do you think so?"* For every set, we coded whether the answers referred to material kind. If the child referred to the material or an according quality at least once (e.g., *"because it is made out of wood," "because it is made out of the same stuff," "because it is just the same"*), the child received one point (i.e., the range of possible scores for the Explanation Task was 0–6).

Conditions: Prediction and explanation task

After the baseline assessment, the four different conditions followed. The materials used in the conditions were composed in a way that was similar to the baseline assessment. That is, we developed six novel object sets following the same logic as described above. However, all objects had different shapes than the objects used in the baseline assessment (and than the objects used in the pretest), and we chose a second standard for each set to be presented in the conditions with two standards (see **Figures 1**, **2** for object sets in the one-standard condition and the two-standard condition, respectively).

The instructions and the number of standards varied for the four conditions (see **Table 1**). The children saw either one or two standards, and the objects of a set were either labeled or unlabeled. In the labeled conditions, the standards were labeled according to their material after their floating behavior had been shown (e.g., *"Look, this one is made out of wood and it floats in water. And this one is also made out of wood and it floats, too"*

TABLE 2 | Object sets in Experiments 1, 2.

	Baseline		Conditions	
Object Set	**Standard**	**Choices**	**Standard**	**Choices**
Wood	Wooden plate	Wooden spoon Metal plate Metal needle Glass marble	Wooden star (wooden ball)	Wooden cube Clay star Clay marble Metal fragment
Clay	Clay fragment	Clay puppet Wax fragment Wax block Cork block	Clay marble (clay stick)	Clay fragment Styrofoam marble Styrofoam stick Cork block
Plastic	Plastic knife	Plastic plate Wooden knife Wooden block Wax fragment	Plastic ruler (plastic plate)	Plastic spoon Wooden ruler Wooden plate Styrofoam ring
Wax	Wax sphere	Wax block Glass sphere Glass marble Metal nut	Wax star (wax fragment)	Wax candle Clay star Clay fragment Glass marble
Metal	Metal ring	Metal marble Wooden ring Wooden button Wax block	Metal spoon (metal ball)	Metal nut Wooden spoon Wooden block Wax candle
Styrofoam	Styrofoam pyramid	Styrofoam block Metal needle Metal nut Clay pyramid	Styrofoam sphere (Styrofoam cube)	Styrofoam plate Glass cube Glass sphere Metal needle

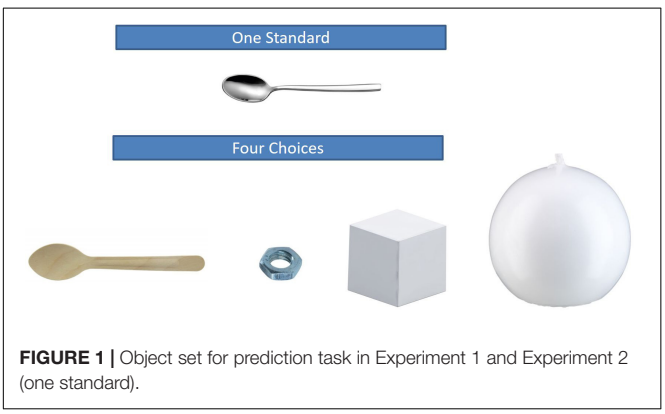


FIGURE 1 | Object set for prediction task in Experiment 1 and Experiment 2 (one standard).

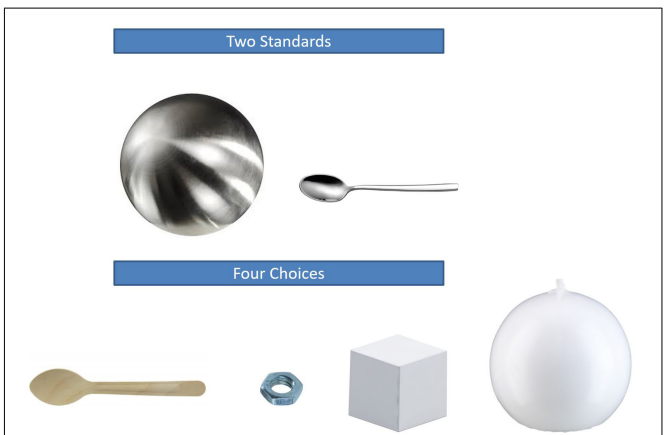


FIGURE 2 | Object set for prediction task in Experiment 1 and Experiment 2 (two standards).

in the two-standard condition). To increase the salience of labels, the labeling was not only applied to the standards but also to all selection objects.

In the Prediction Task, the children were asked to choose one of the selection objects that would show the same floating behavior ("*Which of these also floats in water, just like these two? This one made out of iron, this one made out of wood, this one made out of glass, or this one made out of clay?*"). Afterward, the children were asked to explain their selection (Explanation Task). The Prediction and the Explanation Tasks were scored as described for the baseline assessment (i.e., the range of scores was 0–6). In all conditions, the children were allowed to touch the objects and hold them in their hands, but they were not allowed to put them into the water. Only the experimenter immersed the standards into water for the children to observe. If the children expressed a wish to immerse objects into the water, they were told that they could do so after the experiment was finished. In all phases of the experiment, the children were praised for their active participation; however, no feedback was given concerning the accuracy of their replies.

Results

Table 3 presents the means and standard deviations of the Matching Task and the Labeling Task (pretest), as well as of the Prediction Task and the Explanation Task for the baseline assessment and four conditions. The scores of the Prediction

TABLE 3 | Means (standard deviations) of pretest, prediction task, and explanation task by condition in Experiment 1.

	Pretest		Prediction Task		Explanation Task		
Condition	**Matching**	**Labeling**	**Baseline**	**Experimental**	**Baseline**	**Experimental**	***N***
One_unlabeled	5.60 (1.55)	3.07 (2.05)	2.13 (1.35)	1.73 (1.43)	2.0 (1.56)	1.13 (1.68)	15
One_labeled	6.07 (1.07)	2.78 (1.31)	1.43 (1.22)	3.21 (1.58)	0.72 (0.92)	1.86 (2.14)	14
Two_unlabeled	6.07 (1.39)	2.93 (1.83)	1.73 (1.39)	2.67 (1.59)	0.54 (0.52)	0.80 (1.20)	15
Two_labeled	6.20 (1.57)	3.00 (1.60)	1.93 (0.96)	4.13 (1.81)	1.07 (0.96)	2.80 (2.31)	15

and the Explanation Tasks of the baseline assessment were employed as covariates in the respective task analyses of children's performance in the different conditions.

Before presenting the results with regard to our hypotheses, we present the results of the preliminary analyses testing whether there were significant differences across conditions on the two pretest tasks and the children's performance in the Prediction Task and Explanation Task in the baseline assessment. One-way ANOVAs showed no significant differences across the experimental conditions for the pretests (Matching Task: $p = 0.67$, $\eta_p^2 = 0.028$; Labeling Task: $p = 0.98$, $\eta_p^2 = 0.004$). The Labeling Task (36% correct) was more difficult than the Matching Task (74% correct). There was also no significant difference in the Prediction Task for the baseline assessment across conditions ($p = 0.47$, $\eta_p^2 = 0.04$). There was, however, a significant difference in the Explanation Task of the baseline assessment across conditions, $F(3,55) = 5.68$, $p = 0.002$, $\eta_p^2 = 0.24$. Follow-up analyses of the Explanation Task indicated that children in the one_unlabeled condition had significantly higher scores than children in the one_labeled and the two_unlabeled conditions (see **Table 3**).

To test Hypotheses 1 and 2 (i.e., better performance in the two-standard conditions and an interaction effect of Number of Standards and Labeling), a 2 × 2 ANCOVA with Number of Standards (one vs. two) and Labeling (unlabeled vs. labeled) and the Prediction Task performance as the dependent measure was computed, using the performance in the Prediction Task in the baseline assessment as covariate. The covariate significantly predicted the dependent measure, $F(1,54) = 4.35$, $p = 0.04$, $\eta_p^2 = 07$. As expected, we found a significant effect of the Number of Standards, $F(1,54) = 4.97$, $p = 0.03$, with overall higher mean accuracy in the two-standard condition, $M = 3.40$ ($SD = 1.83$) than in the one-standard condition, $M = 2.45$ ($SD = 1.66$). However, the size of the effect was rather small ($\eta_p^2 = 0.08$). We also found a significant but small main effect of Labeling, $F(1,54) = 4.21$, $p = 0.01$, $\eta_p^2 = 0.04$, with higher mean accuracy in the labeled conditions, $M = 3.69$ ($SD = 1.73$) than in the unlabeled conditions, $M = 2.20$ ($SD = 1.56$). Contrary our expectation, there was no interaction of Number of Standards × Labeling, $p = 0.69$, $\eta_p^2 = 0.003$.

We performed the same analyses with the performance in the Explanation Task as a dependent measure, using performance in the Explanations Task in the baseline assessment as covariate. The covariate significantly predicted this dependent measure, $F(1,54) = 6.26$, $p = 0.015$, $\eta_p^2 = 0.10$. We found only a significant effect of Labeling, $F(1, 54) = 11.00$, $p = 0.002$, $\eta_p^2 = 0.17$, with $M = 2.34$ ($SD = 2.24$) for the labeled conditions and $M = 0.97$ ($SD = 1.45$) for the unlabeled conditions. Contrary to our hypotheses, however, there was no significant main effect of Number of Standards, $p = 0.20$, $\eta_p^2 = 0.03$, and no significant Number of Standards × Labeling interaction, $p = 0.82$, $\eta_p^2 = 0.001$.

In exploratory *post hoc* analyses, we checked whether age contributed to the reported effects. We did not find significant differences between younger children (≤64 months, $N = 36$) and older children (>65 months, $N = 37$) in the Prediction Task, $M = 3.0$ ($SD = 1.85$), $M = 3.13$ ($SD = 1.62$), respectively, $p = 0.79$, $\eta_p^2 = 0.01$, or in the Explanation Task, $M = 1.39$ ($SD = 1.92$), $M = 1.27$ ($SD = 1.90$), respectively, $p = 0.67$, $\eta_p^2 = 0.01$.

Discussion of Experiment 1

In Experiment 1, we investigated whether the presentation of two standards of the same material would support children in relying on the concept of material kind when predicting and explaining objects' floating or sinking, and whether the labeling of objects with their respective materials would increase children's ability to induce the concept of material kind. We found that preschoolers were indeed more likely to base their predictions on the concept of material kind when comparison was triggered by the presentation of two standard objects, with an overall small but significant effect. When presented with two standard objects, children decreased their reliance on irrelevant features, such as shape or size, when predicting floating or sinking. However, our hypothesis that the effect of comparison would be intensified by presenting labels was not confirmed. While we found a main effect of Labeling in the Prediction Task, we did not find a significant interaction between the factors Number of Standards and Labeling. Unexpectedly, even in the one-standard condition, the children were more likely to choose objects with the same material if they were provided with objects' material labels. This finding suggests that children in the one-standard condition used the common label of the standard and the same-material item of the selection objects to derive conclusions with respect to their underlying commonalities.

In the Explanation Task, we found an effect only for Labeling. As for the Prediction Task, the children showed a significantly higher tendency to refer to material when explaining their choices in the conditions when labels were used. In contrast to the results of the Prediction Task and in contrast to our hypothesis, however, there was no effect of the Number of Standards factor. It would appear that the children were able to base their predictions on material kind in conditions with two standards, but they did not explicate their intuitive knowledge when prompted for explanations unless they were supported by

labels. On average, the solution rates for explanations were lower than for predictions. Therefore, one might speculate that this task placed a higher demand on children with regard to the retrieval of conceptual knowledge. A significant group difference had already been detected in the Explanation Task of the baseline assessment prior to the experimental variation, and this *a priori* difference may have diluted the effect of producing explanations in the experimental conditions. We accounted for this difference by using the baseline performance in the Explanation Task as a covariate. Nevertheless, this unexpected baseline difference may have biased performance in the Explanation Task. Thus, these results should be treated with caution.

How may we explain the effect of labeling on children's performance? First, it is likely that the labels presented in Experiment 1 elicited comparison processes that directed the children's attention to the objects' material, especially because material names were used to label all objects. While developmental research has shown that basic-category labels prompt analogical reasoning by comparison in children from an early age (Gelman and Markman, 1986; Davidson and Gelman, 1990; Nguyen and Gelman, 2012; Childers, 2020), Experiment 1 showed that material labels that provide superordinate information may also support inductive reasoning to derive predictions and explanations for floating and sinking. Second, using material labels may activate specific prior conceptual knowledge of material kind. For example, children may know that "wood" refers to the specific quality that some floating objects are made of. Thus, the children may have been more likely to pick wooden objects due to their prior experience. As we applied material labels to the standards and to all selection objects, the children's attention was drawn to material kind as a quality of all the presented objects. Thus, the activation of prior knowledge may have been especially strong. Finally, the effects of Number of Standards and Labeling in the Prediction Task were rather small. In contrast to Namy and Gentner's (2002) study, we used a baseline assessment of children's initial performance. Using this baseline performance as a covariate increases the power to detect effects. Moreover, Namy and Gentner investigated basic-category learning, whereas our study employed tasks in a scientific reasoning context. It is likely that this type of task led to rather small experimental effects since conceptual restructuring in science is difficult to achieve with short-term instructional interventions (Schneider et al., 2012). Specifically, the induction of a concept of material kind may be regarded as a process of initial conceptual restructuring. Since we chose our selection objects in the Prediction and Explanation Tasks on the basis of well-known misconceptions about floating and sinking at preschool and elementary school age (Leuchter et al., 2014), this task requires children to inhibit answers due to misconceptions to make the correct prediction and provide the correct explanation.

To evaluate whether an effect of labeling is indeed due to the activation of prior conceptual knowledge, we contrasted the use of real material labels and made-up labels in Experiment 2. If the labeling effect were due to the activation of prior conceptual knowledge, then we would find effects only for real labels and not for made-up labels To diminish the effects of drawing attention to material kind as a dimension of all objects in a set, we labeled only the standards and not the selection objects.

EXPERIMENT 2

As in Experiment 1, we varied whether one or two standards were presented and whether the standards were unlabeled (i.e., *"this/these"*) or labeled with respect to their material (e.g., *"this/these one/s is/are made of wood"*). In addition to the use of real material labels, we included two conditions in which we used made-up labels (e.g., *"these ones are made of idoform"*). If the labeling effect in Experiment 1 was due to the triggering of analogical reasoning by comparison, then the effect would appear only in the two-standard conditions. If it was due to the triggering of prior conceptual knowledge by the use of the real material labels, then the effect would only appear in the labeled conditions with real material names. In addition to the Prediction and the Explanation Tasks of Experiment 1, we assessed whether there was evidence of conceptual knowledge transfer with regard to predicting whether novel objects would float or sink. We therefore employed a test of conceptual knowledge immediately before the baseline assessment (transfer pretest) and after the children had completed the conditions (transfer posttest).

Our research questions and hypotheses were as follows:

(3) Will preschoolers induce the concept of material kind when analogical reasoning by comparison is triggered by the presentation of two standards with the same floating behavior?
Hypothesis (3): In conditions with two standards, preschoolers will outperform participants in conditions with one standard on the Prediction and the Explanation Tasks.

(4) Does the use of material labels trigger analogical reasoning by comparison?
Hypothesis (4): The use of real material labels will improve performance in the two-standard condition in the Prediction and Explanation Tasks, but not the use of made-up labels.

(5) Will the intervention lead to conceptual knowledge transfer?
Hypothesis (5): There will be knowledge transfer in the two-standard conditions.

Method

Participants

Ninety-nine children from Central Switzerland with German as a first language participated (testing was in German). They were recruited with parental consent through preschools. The mean age was 5 years, 8 months, ranging from 4 years, 4 months to 7 years. The children were mostly from middle-class families living in suburban areas.

Design

Experiment 2 was based on a 2 × 3 between-groups design, including a baseline assessment as in Experiment 1. After the pretest and the baseline assessment, the preschoolers took a conceptual knowledge transfer test (transfer pretest). They then participated in one of six conditions: (1) One Unlabeled Standard (one_unlabeled), (2) One Standard Labeled with Real Material Label (one_real), (3) One Standard Labeled with Made-up

TABLE 4 | Conditions and instructions in Experiment 2.

	One Standard	Two Standards
Unlabeled	Look, this one floats/sinks in water. Which of these also floats/sinks just like this one?	Look, these two float/sink in water. Which of these also floats/sinks just like these two?
Real Label (example: wood)	Look, this one is made out of wood and it floats in water. Which of these also floats just like this one?	Look, this one is made out of wood and it floats in water. And this other one is also made out of wood and it floats, too. Which of these also floats just like these two?
Made-up Label	Look, this one is made out of feb and it floats in water. Which of these also floats just like this one?	Look, this one is made out of feb and it floats in water. And this other one is also made out of feb and it floats, too. Which of these also floats just like these two?

Material Label (one_ made-up), (4) Two Unlabeled Standards (two_unlabeled), (5) Two Standards Labeled with Real Material Label (two_real), or (6) Two Standards Labeled with Made-up Material Label (two_ made-up). After the conditions, the conceptual knowledge transfer test was again presented (transfer posttest). **Table 4** gives an overview of the different conditions and the respective instructions.

Tasks and Procedure

The material, tasks, and procedure employed in Experiment 2 were largely similar to those of Experiment 1. In addition, we employed a conceptual knowledge transfer test that was first conducted immediately before the baseline assessment and then repeated after the children had solved the tasks in the different conditions.

Pretest: Matching/labeling task

The material and procedure for these tasks were the same as in Experiment 1.

Baseline assessment and conditions: prediction and explanation task

The object sets (see **Table 2**) and procedure were the same as in Experiment 1 for the baseline assessment and for the unlabeled conditions (one_unlabeled, two_unlabeled). For the labeled conditions, the procedure differed from Experiment 1 in the following way. In the conditions using real material labels (one_real, two_real), the standards were labeled while their floating behavior was demonstrated (e.g., *"Look, this is made out of wood and it floats in water. And this is also made out of wood and it floats, too."*). Afterward, the children were asked to select one of the selection objects (*"Which of these also floats/sinks in water, just like these two? This one, this one, this one, or this one?"*). Thus, in contrast to the procedure in Experiment 1, we applied the labels only to the standards and not to the four selection objects. In the conditions with made-up labels (one_made-up, two_made-up), the procedure was the same as in the real material label conditions except that the real material labels were replaced with made-up labels. For all objects sets, the children had to choose one of the selection objects (Prediction Task) and to explain their selection (Explanation Task). These two tasks were scored as in Experiment 1.

Test on conceptual knowledge transfer of floating and sinking

We designed a conceptual knowledge test to measure potential knowledge transfer. This test also included Prediction and Explanation Tasks. In the Transfer Prediction Task, the children were asked to predict whether a presented object would float or sink in water. In the Transfer Explanation Task, the children were asked to verbally explain their answers. Five objects were used, all consisting of material that was also used in the object sets for the baseline assessment and in the different conditions. However, the objects in the transfer tasks had novel shapes and were selected because they represented common misconceptions based on the perceptual qualities of size, weight, or shape: A large and heavy wooden block, a thin wooden board with holes, a metal cube, a small metal needle, and a large block of Styrofoam (see also Leuchter et al., 2014, for a similar task). When predicting whether such objects float or sink in water, children typically refer to their size, weight, or shape in their explanations rather than to their material. Each object was first shown to the child, and the child was asked to touch it and to hold it. A container with water was placed onto the table next to the objects but the children were only allowed to put the objects into it after all tasks were finished. In the Transfer Prediction Task, the child was first asked *"Does this float or sink in water?"* In the immediately following Transfer Explanation Task, the child was asked *"Why do you think so?"* In the Transfer Prediction Task, the children received one point for a correct prediction (i.e., range of scores 0 – 5). In the Transfer Explanation Task, the children received one point if they provided a correct explanation with regard to material kind (i.e., range of scores 0–5).

Results

Table 5 presents the means and standard deviations of the six conditions for the pretest scores (Matching Task, Labeling Task), the Prediction Task, and the Explanation Task in the baseline assessment and the different conditions. **Table 6** presents the respective descriptive values for the pre- and posttest scores of the Transfer Prediction Task and the Transfer Explanation Task.

In preliminary analyses, we checked whether the children differed among conditions with respect to their pretest and baseline assessment performances using one-way ANOVAs. There were no significant group differences (Matching Task: $p = 0.16$, $\eta_p^2 = 0.08$; Labeling Task: $p = 0.56$, $\eta_p^2 = 0.04$; baseline Prediction Task: $p = 0.64$, $\eta_p^2 = 0.035$; baseline Explanation Task: $p = 0.89$, $\eta_p^2 = 0.018$; pretest Transfer Prediction Task: $p = 0.36$, $\eta_p^2 = 0.056$; pretest Transfer Explanation Task: $p = 0.64$, $\eta_p^2 = 0.035$).

To test our hypotheses derived from research questions 3 and 4 (i.e., better performance in the two-standard conditions and

TABLE 5 | Means (standard deviations) of pretest, prediction task, and explanation task by condition in Experiment 2.

	Pretest		Prediction Task		Explanation Task		
Condition	**Matching**	**Labeling**	**Baseline**	**Experimental**	**Baseline**	**Experimental**	***N***
One_unlabeled	5.65 (1.79)	1.78 (1.21)	2.56 (1.54)	2.67 (2.12)	1.55 (1.95)	1.55 (1.85)	18
One_real	4.56 (2.00)	2.17 (2.12)	2.72 (1.81)	3.39 (2.15)	1.72 (2.27)	2.11 (2.35)	18
One_made-up	5.73 (2.22)	2.73 (1.62)	1.87 (2.03)	2.13 (2.23)	1.53 (2.09)	2.27 (2.49)	15
Two_unlabeled	6.11 (1.90)	2.67 (1.85)	2.89 (1.60)	3.94 (2.41)	2.22 (2.12)	3.33 (2.50)	18
Two_real	6.18 (1.13)	2.41 (1.18)	2.24 (1.79)	3.65 (2.09)	1.41 (1.94)	2.53 (2.53)	17
Two_made-up	5.46 (2.50)	2.54 (1.90)	2.54 (2.07)	3.77 (2.13)	1.69 (1.93)	2.85 (2.48)	13

TABLE 6 | Means (standard deviations) of conceptual knowledge transfer by condition in Experiment 2.

	Transfer Prediction Task		Transfer Explanation Task		
Condition	**Pretest**	**Posttest**	**Pretest**	**Posttest**	***N***
One_unlabeled	3.11 (0.96)	3.39 (1.29)	0.67 (1.33)	1.05 (1.43)	18
One_real	3.61 (0.85)	3.72 (0.75)	0.94 (1.39)	1.11 (1.45)	18
One_made-up	3.13 (0.91)	3.07 (1.10)	0.53 (1.06)	0.53 (0.91)	15
Two_unlabeled	3.39 (1.04)	3.61 (0.85)	0.67 (1.19)	1.44 (1.69)	18
Two_real	3.00 (1.06)	3.05 (0.97)	0.41 (1.00)	0.65 (0.86)	17
Two_made-up	3.54 (0.97)	3.46 (1.27)	1.08 (1.04)	1.54 (2.02)	13

an improvement in performance in the two-standard condition with real labels), we conducted a 2 × 3 ANCOVA with Number of Standards (one vs. two) and Labeling (no material label vs. real material label vs. made-up label) as between-subject factors, and performance in the Prediction Task in different conditions as the dependent measures, including the baseline Prediction Task performance as covariate. The covariate predicted performance in the Prediction Task significantly, $F(1,92) = 92.77$, $p = 0.001$, $\eta_p^2 = 0.50$. As expected, we found a main effect for Number of Standards, $F(1,92) = 8.28$, $p = 0.005$, with higher means in the conditions with two standards, $M = 3.79$ ($SD = 2.18$) than in the conditions with one standard, $M = 2.76$ ($SD = 2.18$). However, the effect size was rather small ($\eta_p^2 = 0.08$). However, we did not find an effect of Labeling, $p = 0.49$, $\eta_p^2 = 0.01$, nor an interaction of Number of Standards × Labeling, $p = 0.87$, $\eta_p^2 = 0.003$. Overall, the results suggest that presenting two objects of the same material benefits children's induction of the concept of material kind. Labeling the material of the standards did not cause significant performance differences, regardless of whether real or made-up labels were used or whether labels were combined with the presentation of two objects.

For the Explanation Task performance, we conducted the same 2 × 3 ANCOVA as for the Prediction Task. We found a similar pattern. There was a significant effect for Number of Standards, $F(1,92) = 6.16$, $p = 0.015$, $\eta_p^2 = 0.063$, with $M = 2.92$ ($SD = 2.47$) for the two-standard conditions and $M = 1.96$ ($SD = 2.21$) for one-standard conditions. This finding suggests that preschoolers were more likely to explain their choices by reference to material kind when two standards were presented than when only one standard was given. There was no effect for Labeling ($p = 0.66$, $\eta_p^2 = 0.01$) and no Number of Standards × Labeling interaction ($p = 0.60$, $\eta_p^2 = 0.011$). The covariate predicted performance in the Explanation Task significantly, $F(1,92) = 127.67$, $p < 0.001$, $\eta_p^2 = 0.58$. In additional exploratory analyses, we found no significant differences between the age groups of children younger than 60 months ($N = 11$), between 61 and 72 months ($N = 51$), and more than 72 months ($N = 20$) in analyses of covariance with the Prediction Task, $M = 2.91$ ($SD = 2.43$), $M = 3.12$ ($SD = 2.12$), $M = 3.95$ ($SD = 2.23$), respectively, $p = 0.49$, $\eta_p^2 = 0.02$, or the Explanation Task, $M = 0.45$ ($SD = 0.82$), $M = 0.94$ ($SD = 1.45$), $M = 1.45$ ($SD = 1.57$), respectively, $p = 0.65$, $\eta_p^2 = 0.01$.

Finally, to test research question 5 (i.e., whether the conditions increased conceptual knowledge transfer), we tested performance in the Transfer Prediction and Explanation Tasks. First, we conducted a 2 × 2 × 3 repeated measurement ANOVA with the pretest and posttest scores for the Transfer Prediction Task as within-subjects factors and Number of Standards (one vs. two) and Labeling (unlabeled vs. real material label vs. made-up label) as between-subjects factors. Contrary to our expectations, we did not find a significant interaction for Time × Number of Standards, $p = 0.85$, $\eta_p^2 = 0.000$, nor main effects for Time, $p = 0.39$, $\eta_p^2 = 0.008$, Number of Standards, $p = 0.98$, $\eta_p^2 = 0.000$, Labeling, $p = 0.94$, $\eta_p^2 = 0.001$, or interaction effects for Time × Labeling, $p = 0.45$, $\eta_p^2 = 0.017$, or Time × Labeling × Number of Standards, $p = 0.99$, $\eta_p^2 = 0.000$. Second, we conducted a similar 2 × 2 × 3 repeated measurement ANOVA with the pretest and posttest scores for the Transfer Explanation Task. This analysis showed a significant main effect of Time, $F(1,93) = 9.43$, $p < 0.01$, $\eta_p^2 = 0.09$, indicating an improvement from the pretest, $M = 0.71$ ($SD = 1.18$) to the posttest, $M = 1.05$ ($SD = 1.44$). There were no significant main effects for Number of Standards, $p = 0.52$, $\eta_p^2 = 0.004$, or Labeling, $p = 0.81$, $\eta_p^2 = 0.005$, nor interaction effects for Time × Number of Standards, $p = 0.17$, $\eta_p^2 = 0.02$, Time × Labeling, $p = 0.27$, $\eta_p^2 = 0.028$, or

Time × Labeling × Number of Standards, $p = 0.74$, $\eta_p^2 = 0.006$. That is, the improvement from pretest to posttest in the Explanation Task did not differ between conditions.

Discussion of Experiment 2

As expected, we found that preschoolers were more likely to induce the concept of material kind when analogical reasoning by comparison was triggered by presenting two objects of the same material rather than only one object. This effect was reflected in the Prediction Task and the Explanation Task, albeit with small effect sizes overall. As in Experiment 1, children's performance in the baseline assessment, included as a covariate in our statistical models, contributed significantly to children's performance in the Prediction and Explanation Tasks, with a large effect size. In Experiment 1, we also found that assigning a common label to all objects increased choices of objects of the same material. In contrast to Experiment 1, we used the material labels in a more restricted way in Experiment 2, applying them only to the standards and not to the selection objects. This more restricted use did not improve preschoolers' performance in the Prediction and Explanation Tasks in comparison to the unlabeled conditions. Our analyses of the Transfer Prediction Task and the Transfer Explanation Task only revealed a small overall gain in the Transfer Explanation Task. Overall, none of the conditions caused specific knowledge transfer effects. However, do our results suggest that the use of labels promotes analogical reasoning by comparison? The lack of a difference between the "two standards/no label" and "two standards/real label" conditions suggests that using material labels has no effect on top of presenting two objects. Given that the instruction and observation of the floating and sinking of two objects might have already invoked the comparison of the two standards, children may not have needed an additional linguistic prompt to align both items.

GENERAL DISCUSSION

Task Effects of Comparisons of Number of Standards and Linguistic Labels

In our two experiments, we investigated whether triggering analogical reasoning by comparison and additional labeling would enhance preschoolers' induction of the concept of material kind as a basis for the generation of predictions and explanations in a scientific context for "floating and sinking." To this end, preschoolers were randomly assigned to conditions that were intended to trigger analogical reasoning by comparison by presenting objects evidenced to sink or float in water (i.e., standards). The objects were presented without labels, with their real material labels, or with made-up labels (Experiment 2 only). The children were then asked to predict which of four selection objects with an unknown status would float or sink and to explain their prediction. As expected, the presentation of two standards rather than one standard supported preschoolers' induction of the concept of material kind as a basis for generating predictions in both experiments (Ruffman et al., 1993). The benefit of two standards for the provision of explanations only emerged in the second experiment, however. This pattern of findings fits with previous research emphasizing the beneficial role for category and concept learning of triggering analogical reasoning by comparison through the presentation of two standards (Loewenstein et al., 1999; Gentner et al., 2007; Gentner, 2010, 2016). Unlike previous research on preschoolers' learning through comparison, which employed forced-choice tasks, the present experiments assessed children's induction of conceptual knowledge for hypothesis generation using tasks that required children to select an object (Prediction Task) and to explain their selection (Explanation Task).

Prior research on analogical reasoning by comparison has provided some evidence that labels as language prompts may be even more beneficial than simply juxtaposing entities or objects. Language is presumed to play a pivotal role because using common labels across different entities may function as an invitation to compare and, as such, to align the similarities and differences of the entities and the encoding and abstraction of a generalizable schema (Gentner and Namy, 1999, 2006; Namy and Gentner, 2002; Gentner, 2010, 2016). In Experiments 1 and 2, we labeled the material comprising each object, potentially providing superordinate category labels. Contrary to our hypothesis, we found an effect of labeling both in the conditions with one standard and with two standards. Since the labels in Experiment 1 were not only applied to the presented standards but also to the four selection objects, it is likely that the children's prior conceptual knowledge of material kind was activated in the condition with a single standard, facilitating responses based on the same material. The use of real material labels in Experiment 1 differed from the approach of Namy and Gentner (2002), who used made-up labels. In Experiment 1, at least some of the children were familiar with material labels from everyday life contexts, as indicated by their performance in the pretest, in which the children were asked to label the material of various objects. In Experiment 2, we therefore used labels in a more restricted way in order to differentiate between the effects of prior conceptual knowledge and the facilitation of comparison by labeling. We found no effect for labeling with this more restricted use. Children's predictions and explanations did not differ from the conditions without labels, neither when made-up labels were used nor when real labels were used. In Gentner and Namy's (1999) Experiment 2, a label and a no-label condition were contrasted, as well as a compare (two standards) and a non-compare (one standard) condition. Gentner and Namy found that the label/two-standard standard condition significantly increased appropriate responses not only in contrast to both single standard conditions but also to the unlabeled/two-standard condition. How may these differences be explained? We suppose that even our "two standards without labeling" condition sufficed as an invitation for analogical reasoning by comparison because the children could also observe whether these objects floated or sank as empirical evidence associated with the respective objects. These observations may have created an alignable similarity between the two objects, helping the preschoolers to induce the concept of material kind, in both Experiment 1 and Experiment 2. By contrast, Namy and Gentner only showed pictures of one versus two objects; thus, verbal support may have been more important than in our experiments. Moreover, in our experiments, all the standard objects of the same material showed

the same floating/sinking behavior. That is, the covariation between objects and observed object behavior was perfect. In this regard, our findings fit with previous research indicating that even 4-year-olds can coordinate theory and evidence if there is a perfect covariation (Tullos and Woolley, 2009).

In Experiment 2, preschoolers' performance in the conceptual knowledge transfer test did not differ between conditions. Even so, the overall significant gains from pretest to posttest on the dependent measure of explanations may provide a first, albeit weak indication that a process of conceptual restructuring was initiated. Because conceptual restructuring is a process in which naïve conceptions are gradually extended and refined (Vosniadou et al., 2001), this process needs to be continued and deepened since children typically need additional opportunities to overcome their misconceptions (Leuchter et al., 2014). The slow process of conceptual restructuring may explain why the benefits of task environments inviting analogical reasoning by comparison are rather small in our experiments. Our tasks required preschoolers to overcome their misconceptions and construct novel conceptual knowledge in a science domain with a high degree of prior knowledge—the concept of material kind within the context of floating and sinking. Developing such knowledge may be more challenging than the basic-category learning studied in previous research on the benefits of comparison (e.g., Namy and Gentner, 2002).

Limitations

Throughout this paper, we suggested that analogical reasoning by comparison benefits young children's generation of predictions and explanations as a central aspect of scientific reasoning. The cognitive models of analogical reasoning by comparison provide conceptualizations of how children actually engage in this process: When learners engage in analogical reasoning by comparison, they identify similarities and differences between entities, they may align them and abstract a schema, they may re-represent their existing conceptions, and they may project inferences based on the results of these processes (Gentner, 2010). In our experiments, we investigated only the induction of one specific basic science concept: the concept of material kind. We did not investigate hypothesis generation in other science domains that vary in children's prior conceptual knowledge and that may have an impact on children's generation of predictions and explanations. Therefore, further research is needed to investigate the induction of other scientific concepts so that the claim that analogical reasoning by comparison has benefits for scientific reasoning in general may be strengthened. Moreover, while we did assess children's prior knowledge of material kind, we did not look at individual differences and their influences on the process of hypothesis generation. Thus, further research should address the interaction of prior knowledge and the generation of predictions and explanations. Generally, while hypothesis generation is a central aspect of scientific reasoning, it will be important for future research to investigate whether other elements of the scientific inquiry cycle, such as the generation of experiments, the interpretation of data patterns, and the evaluation of evidence, would also benefit from interventions aimed to trigger analogical reasoning by comparison.

While both our studies were conducted with preschoolers, we included a wide age range of children, between 4 and 7 years. Additional exploratory analyses indicated that there were no significant differences between different age groups. However, our study design does not make it possible to fully disentangle the possible impact of age on the tasks used in the present experiments. On the basis of results summarized by Gentner (2010) and Hespos et al. (2020), we expect analogical reasoning by comparison to be a cognitive mechanism available to children starting at an early age. However, further research is needed to investigate age differences with the task formats that we employed in our experiments. For example, it is possible that age may be a more important factor in the Explanation Task, which required children to produce verbal answers, than in the Prediction Task, even though we did not find effects from age in either task format. Here, we have to acknowledge that the power of the present experiments is not sufficient to detect such differences (i.e., interactions between age, tasks, and conditions). A closer look at age differences may also illuminate the differences in performance between the dependent measures of Prediction and Explanation Tasks found in Experiments 1 and 2. Therefore, task designs taking into account individual differences in children's responses could also contribute to understanding the contingencies between children's predictions and explanations.

Conclusion

In our two experiments, we investigated whether triggering analogical reasoning by comparison would benefit children's predictions and explanations of objects' floating or sinking based on the concept of material kind. On the one hand, we triggered analogical reasoning by comparison by presenting two objects of the same material simultaneously and found some evidence that this indeed benefited children's induction of the concept of material kind for hypothesis generation. On the other hand, unexpectedly, the additional provision of language prompts did not increase the effect of presenting two objects. In previous research, such additional support was often necessary for concept learning, especially with young children. It may be that our task structure provided a different kind of support for triggering comparison because our tasks included the demonstration of the floating and sinking behavior of the standards. This speculation may provide interesting directions for future research. Our experiments therefore only give a first hint that analogical reasoning by comparison may be helpful for the induction of science concepts in a scientific reasoning context.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation

and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

HS and IH contributed the conception and design of the experiments. ML, HS, IH, and LS conducted the Experiments 1 and 2 in Germany and Switzerland. HS and IH performed the statistical analyses. ML, HS, and IH wrote a first draft of the manuscript. All the authors contributed to manuscript revision and approved the submitted version.

FUNDING

This research was funded with internal research funds from the authors' institutions as well as internal funding from the Max Planck Institute for Human Development, Berlin, Germany.

REFERENCES

Alfieri, L., Nokes-Malach, T. J., and Schunn, C. D. (2013). Learning through case comparisons: a meta-analytic review. *Educ. Psychol.* 48, 87–113. doi: 10.1080/00461520.2013.775712

Arunachalam, S., and Waxman, S. R. (2010). Language and conceptual development. *Wiley Interdiscip. Rev. Cogn. Sci.* 1, 548–558. doi: 10.1002/wcs.37

Carey, S. (2000). Science education as conceptual change. *J. Appl. Dev. Psychol.* 21, 13–19.

Childers, J. (2020). *Language and Content Acquisition from Infancy through Childhood: Learning from Multiple Exemplars*. Berlin: Springer.

Chinn, C. A., and Brewer, W. F. (2001). Models of data: a theory of how people evaluate data. *Cogn. Instr.* 19, 323–393. doi: 10.1207/S1532690XCI1903_3

Chinn, C. A., and Malhotra, B. A. (2002). Children's responses to anomalous scientific data: how is conceptual change impeded? *J. Educ. Psychol.* 94, 327. doi: 10.1037//0022-0663.94.2.327

Christie, S. (2020). "Multiple exemplars of relations," in *Language and Content Acquisition from Infancy through Childhood: Learning from Multiple Exemplars*, ed. J. Childers (Berlin: Springer), 221–246. doi: 10.1007/978-3-030-35594-4_11

Christie, S., and Gentner, D. (2010). Where hypotheses come from: learning new relations by structural alignment. *J. Cogn. Dev.* 11, 356–373. doi: 10.1080/15248371003700015

Davidson, N. S., and Gelman, S. A. (1990). Inductions from novel categories: the role of language and conceptual structure. *Cog. Dev.* 5, 151–176. doi: 10.1016/0885-2014(90)90024-n

Dickinson, D. K. (1987). The development of a concept of material kind. *Sci. Educ.* 71, 615–628. doi: 10.1002/sce.3730710412

Driver, R., Newton, P., and Osborne, J. (2009). Establishing the norms of scientific argumentation in classrooms. *Sci. Educ.* 84, 1–312. doi: 10.1007/s10503-017-9424-z

Dunbar, K., and Klahr, D. (2012). "Scientific thinking and reasoning," in *The Oxford Handbook of Thinking and Reasoning*, eds K. J. Holyoak and R. G. Morrison (Oxford: Oxford University Press), 701–718.

Gelman, S. A., Collman, P., and Maccoby, E. E. (1986). Inferring properties from categories versus inferring categories from properties: the case of gender. *Child Dev.* 57, 396–404. doi: 10.2307/1130595

Gelman, S. A., and Markman, E. M. (1986). Categories and induction in young children. *Cognition* 23, 183–209. doi: 10.1016/0010-0277(86)90034-x

Gentner, D. (2010). Bootstrapping the mind: analogical processes and symbol systems. *Cogn. Sci.* 34, 752–775. doi: 10.1111/j.1551-6709.2010.01114.x

Gentner, D. (2016). Language as cognitive tool kit: how language supports relational thought. *Am. Psychol.* 71, 650–657. doi: 10.1037/amp0000082

Gentner, D., and Hoyos, C. (2017). Analogy and abstraction. *Top. Cogn. Sci.* 9, 672–693. doi: 10.1111/tops.12278

Gentner, D., Loewenstein, J., and Hung, B. (2007). Comparison facilitates children's learning of names for parts. *J. Cogn. Dev.* 8, 285–307. doi: 10.1080/15248370701446434

Gentner, D., and Medina, J. (1998). Similarity and the development of rules. *Cognition* 65, 263–297. doi: 10.1016/s0010-0277(98)00002-x

Gentner, D., and Namy, L. L. (1999). Comparison in the development of categories. *Cog. Dev.* 14, 487–513. doi: 10.1016/s0885-2014(99)00016-7

Gentner, D., and Namy, L. L. (2006). Analogical processes in language learning. *Curr. Dir. Psychol. Sci.* 15, 297–301. doi: 10.1111/j.1467-8721.2006.00456.x

Gentner, D., and Rattermann, M. (1991). "Language and the career of similarity," in *Perspectives on Language and Thought: Interrelations in Development*, eds S. Gelman and J. Byrnes (Cambridge: Cambridge University Press), 225–277. doi: 10.1017/cbo9780511983689.008

Gentner, D., and Smith, L. (2012). "Analogical reasoning," in *Encyclopedia of Human Behavior* 2nd Ed. V. S. Ramachandran (Oxford: Elsevier), 130–136.

Graham, S. A., Booth, A. E., and Waxman, S. R. (2012). Words are not merely features: only consistently applied nouns guide 4-year-olds' inferences about object categories. *Lang. Learn. Dev.* 8, 136–145. doi: 10.1080/15475441.2011.599304

Gropen, J., Kook, J. F., Hoisington, C., and Clark-Chiarelli, N. (2017). Foundations of science literacy: efficacy of a preschool professional development program in science on classroom instruction, teachers' pedagogical content knowledge, and children's observations and predictions. *Early Educ. Dev.* 28, 607–631. doi: 10.1080/10409289.2017.1279527

Hardy, I., Jonen, A., Möller, K., and Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking.". *J. Educ. Psychol.* 98, 307–326. doi: 10.1037/0022-0663.98.2.30

Hespos, S., Anderson, E., and Gentner, D. (2020). "Structure mapping processes enable infants' learning across domains including language," in *Language and Content Acquisition from Infancy through Childhood: Learning from Multiple Exemplars*, ed. J. Childers (Berlin: Springer), 97–104. doi: 10.1080/01690960344000143

Holland, J. H., Holyoak, K. J., Nisbett, R. E., and Thagard, P. R. (1989). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge: MIT Press.

Holyoak, K. J. (2005). "Analogy," in *The Cambridge Handbook of Thinking and Reasoning*, eds K. J. Holyoak and R. G. Morrison (Cambridge: Cambridge University Press), 117–142.

Johanson, M., and Papafragou, A. (2016). The influence of labels and facts on children's and adults' categorization. *J. Exp. Child Psychol.* 144, 130–151. doi: 10.1016/j.jecp.2015.11.010

Klahr, D. (2000). *Exploring Science: The Cognition and Development of Discovery Processes*. Cambridge, MA: MIT Press.

Klahr, D., and Dunbar, K. (1988). Dual space search during scientific reasoning. *Cogn. Sci.* 12, 1–48. doi: 10.1207/s15516709cog1201_1

Koerber, S., Sodian, B., Thoermer, C., and Nett, U. (2005). Scientific reasoning in young children: preschoolers' ability to evaluate covariation evidence. *Swiss J. Psychol.* 64, 141–152. doi: 10.1024/1421-0185.64.3.141

Köksal-Tuncer, Ö, and Sodian, B. (2018). The development of scientific reasoning: hypothesis testing and argumentation from evidence in young children. *Cog. Dev.* 48, 135–145. doi: 10.1016/j.cogdev.2018.06.011

Kuhn, D. (2010). "What is scientific thinking and how does it develop?," in *Handbook of Childhood Cognitive Development*, 2nd Edn. ed. U. Goswami (Oxford: Blackwell), 472–534.

Kurtz, K. J., and Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: when two problems are better than one. *Mem. Cognit.* 35, 334–341. doi: 10.3758/bf03193454

Kurtz, K. J., Miao, C. H., and Gentner, D. (2001). Learning by analogical bootstrapping. *J. Learn. Sci.* 10, 417–446. doi: 10.1207/s15327809jls1004new_2

Leuchter, M., Saalbach, H., and Hardy, I. (2014). Designing science learning in the first years of schooling: an intervention study with sequenced learning material on the topic of "floating and sinking.". *Int. J. Sci. Educ.* 36, 1751–1771. doi: 10.1080/09500693.2013.878482

Loewenstein, J., and Gentner, D. (2001). Spatial mapping in preschoolers: close comparisons facilitate far mappings. *J. Cogn. Dev.* 2, 189–219. doi: 10.1207/s15327647jcd0202_4

Loewenstein, J., Thompson, L., and Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychon. Bull. Rev.* 6, 586–559. doi: 10.3758/bf03212967

Mercier, H. (2011). Reasoning serves argumentation in children. *Cog. Dev.* 26, 177–191. doi: 10.1016/j.cogdev.2010.12.001

Mercier, H., and Sperber, D. (2009). "Intuitive and reflective inferences," in *Two Minds: Dual Processes and Beyond*, eds J. Evans and K. Frankish (Oxford: Oxford University Press), 149–170. doi: 10.1093/acprof:oso/9780199230167.003.0007

Morris, B. J., Masnick, A. M., Zimmerman, C., and Croker, S. (2012). "The emergence of scientific reasoning," in *Current Topics in Children's Learning and Cognition*, eds H. Kloos, B. Morris, and J. Amaral (London: IntechOpen), 61–82. doi: 10.5772/53885

Namy, L. L., and Gentner, D. (2002). Making a silk purse out of two sow's ears: young children's use of comparison in category learning. *J. Exp. Psychol. Gen.* 131, 5–15. doi: 10.1037//O096-LWS.131.I.5

Nguyen, S. P., and Gelman, A. (2012). Generic language facilitates children's cross-classification. *Cog. Dev.* 27, 154–167. doi: 10.1016/j.cogdev.2012.01.001

Penner, D., and Klahr, D. (1996). The interaction of domain-specific knowledge and domain general discovery strategies: a study with sinking objects. *Child Dev.* 67, 2709–2727. doi: 10.2307/1131748

Piekny, J., Grube, D., and Maehler, C. (2014). The development of experimentation and evidence evaluation skills at preschool age. *Int. J. Sci. Educ.* 36, 334–354. doi: 10.1080/09500693.2013.776192

Piekny, J., and Maehler, C. (2013). Scientific reasoning in early and middle childhood: the development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *Br. J. Dev. Psychol.* 31, 153–179. doi: 10.1111/j.2044835X.2012.02082.x

Ruffman, T., Perner, J., Olson, D. R., and Doherty, M. (1993). Reflecting on scientific thinking: children's understanding of the hypothesis–evidence relation. *Child Dev.* 64, 1617–1636. doi: 10.2307/1131459

Saalbach, H., Imai, M., and Schalk, L. (2012). Grammatical gender and inferences about biological properties in german-speaking children. *Cogn. Sci.* 36, 1251–1267. doi: 10.1111/j.1551-6709.2012.01251.x

Sandoval, W., Sodian, B., Koerber, S., and Wong, J. (2014). Developing children's early competencies to engage with science. *Educ. Psychol.* 49, 139–152. doi: 10.1080/00461520.2014.917589

Schalk, L., Saalbach, H., and Stern, E. (2016). Approaches to foster transfer of formal principles: which route to take? *PLoS One* 11:e0148787. doi: 10.1371/journal.pone.0148787

Schneider, M., Vamvakoussi, X., and van Dooren, W. (2012). "Conceptual change," in *Encyclopedia of the Sciences of Learning*, ed. N. Seel (New York: Springer), 735–738.

Schulz, L. E., Gopnik, A., and Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Dev. Sci.* 10, 322–332. doi: 10.1111/j.1467-7687.2007.00587.x

Smith, C., Carey, S., and Wiser, M. (1985). On differentiation: a case study of the development of the concepts of size, weight, and density. *Cognition* 21, 177–237. doi: 10.1016/0010-0277(85)90025-3

Sodian, B., Zaitchik, D., and Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Dev.* 62, 753–766. doi: 10.2307/1131175

Tullos, A., and Woolley, J. (2009). The development of children's ability to use evidence to infer reality status. *Child Dev.* 80, 101–114. doi: 10.1111/j.1467-8624.2008.01248.x

van der Graaf, J., Segers, E., and Verhoeven, L. (2016). Scientific reasoning in kindergarten: cognitive factors in experimentation and evidence evaluation. *Learn. Individ. Differ.* 49, 190–200. doi: 10.1016/j.lindif.2016.06.006

Vosniadou, S., Ioannides, C., Dimitrakopoulou, A., and Papademetriou, E. (2001). Designing learning environments to promote conceptual change in science. *Learn. Instr.* 11, 381–419. doi: 10.1016/S0959-4752(00)00038-4

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Dev. Rev.* 27, 172–223. doi: 10.1016/j.dr.2006.12.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Hardy, Saalbach, Leuchter and Schalk. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 05 August 2020
doi: 10.3389/fpsyg.2020.01934

Investigating Science Together: Inquiry-Based Training Promotes Scientific Conversations in Parent-Child Interactions

Ian L. Chandler-Campbell[1], Kathryn A. Leech[2] and Kathleen H. Corriveau[1]*

[1] Wheelock College of Education and Applied Human Development, Boston University, Boston, MA, United States, [2] School of Education, The University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

This study examined the effects of two pedagogical training approaches on parent-child dyads' discussion of scientific content in an informal museum setting. Forty-seven children (mean age = 5.43) and their parents were randomly assigned to training conditions where an experimenter modeled one of two different pedagogical approaches when interacting with the child and a science-based activity: (1) a scientific inquiry-based process or (2) a scientific statement-sharing method. Both approaches provided the same information about scientific mechanisms but differed in the process through which that content was delivered. Immediately following the training, parents were invited to model the same approach with their child with a novel science-based activity. Results indicated significant differences in the process through which parents prompted discussion of the targeted information content: when talking about causal scientific concepts, parents in the scientific inquiry condition were significantly more likely to pose questions to their child than parents in the scientific statements condition. Moreover, children in the scientific inquiry condition were marginally more responsive to parental causal talk and provided significantly more scientific content in response. These findings provide initial evidence that training parents to guide their children using scientific inquiry-based approaches in informal learning settings can encourage children to participate in more joint scientific conversations.

Keywords: parental guidance, scientific learning, informal learning, museum learning, science education, pedagogical approaches

OPEN ACCESS

Edited by:
Amanda C. Brandone, Lehigh University, United States

Reviewed by:
Xunyi Lin, Fujian Normal University, China
Natalie Ann Munro, The University of Sydney, Australia
Maureen Callanan, University of California, Santa Cruz, United States

****Correspondence:***
Ian L. Chandler-Campbell
ilc@bu.edu

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *16 February 2020*
Accepted: *13 July 2020*
Published: *05 August 2020*

Citation:
Chandler-Campbell IL, Leech KA and Corriveau KH (2020) Investigating Science Together: Inquiry-Based Training Promotes Scientific Conversations in Parent-Child Interactions. Front. Psychol. 11:1934. doi: 10.3389/fpsyg.2020.01934

INTRODUCTION

Parents, as some of children's first learning partners, play a vital role in scaffolding children's learning about scientific concepts (e.g., Crowley et al., 2001; Vlach and Noll, 2016; Legare et al., 2017). Through informal interaction, parents expose children to scientific content through toys and activities (Jacobs and Bleeker, 2004). Moreover, parents play an active role in fostering children's engagement in science by modeling interest through the questions they pose to children as well as providing explanations to their children's questions (Crowley and Callanan, 1998; Dotterer et al., 2009; Wang and Degol, 2013; Willard et al., 2019). Such explanatory talk can be especially

important, as it supplies children with relevant information and can provide insight into underlying causal mechanisms that children would be unlikely to acquire through first-hand exploration (Crowley and Callanan, 1998; Shtulman and Checa, 2012; Haden et al., 2014; Vlach and Noll, 2016). Indeed, explanatory conversations between parents and their children in informal learning settings can be beneficial for scientific learning outcomes in both the short-term (Haden, 2010; Leichtman et al., 2017) and long-term (Tenenbaum et al., 2005). In the current study, we explored the impact of a brief training session on parent-child conversation in informal science learning for 4- to 6-years-old children. We target two dimensions of effective explanatory talk: the scientific *content* and the process by which it is *delivered* to the child. Below, we expand upon the importance of parent-child conversations about science before turning to our rationale for the current study.

The Role of Parental Explanations in Early Science Learning

Although many parents understand the importance of communicating with their children about science, they vary significantly in their tendency to provide accurate, developmentally-appropriate explanations (e.g., Shtulman and Checa, 2012). Observational research in museum contexts has found that parents often provide brief, incomplete explanations when attempting to communicate scientific information (Crowley et al., 2001). There may be many reasons that parents sometimes provide incomplete explanations, ranging from parents attempting to translate complicated science concepts into developmentally-appropriate explanations or simply being judicious with their time in response to the large number of child-initiated questions (Chouinard et al., 2007; Kurkul and Corriveau, 2018; Yu et al., 2019). Another likely reason for providing incomplete explanations is that parents feel they do not have the knowledge needed to provide scientifically-accurate explanations of scientific phenomena (Crowley et al., 2001; Shtulman and Checa, 2012). Consistent with this hypothesis, evidence from multiple research studies suggests many adults, including parents, lack an accurate understanding of the content associated with many scientific domains (Jipson and Callanan, 2003; Rigney and Callanan, 2011; Shtulman and Checa, 2012; Vlach and Noll, 2016). In some cases, such incomplete or inadequate knowledge may be associated with scientific misconceptions that can then be transferred to children. For example, Shtulman and Checa (2012) found that parents often provided explanations that propagated misconceptions about the causal mechanisms of evolutionary processes, such as the inaccurate idea that animals that share similar physical features are likely to share evolutionary ancestry (see also Kelemen and Rosset, 2009; Coley and Tanner, 2012, 2015).

Parents' possession of – or access to – accurate scientific knowledge is not a panacea. Even when adults are knowledgeable about an underlying scientific mechanism, they are often unsure how to generate a developmentally-appropriate explanation (Vlach and Noll, 2016). In the current study, we focus not only on the *content* young children hear but also on the *delivery* of such content. Our approach is based on a growing literature from both psychology and education highlighting the important role of developing "scientific habits of mind" in early science learning settings (National Research Council [NRC], 2012). For example, the *Framework for K–12 Science Education* highlights the importance of children learning about the iterative process of science and engaging in scientific activities and thinking, such as the practices of "inquiry and investigation, collection and analysis of evidence, logical reasoning, and communication and application of information" (National Research Council [NRC], 2012, p. 250). Note that focusing on the role of parent-child communication does not negate the importance of learning about and participating in the process of scientific experimentation. As noted by Dimension 1 of the *Next Generation Science Standards*, children "cannot comprehend scientific practices, nor fully appreciate the nature of scientific knowledge itself, without directly experiencing those practices for themselves" (NGSS, 2013, p. 5). Instead, our approach is drawn from social constructivist models of learning where the dyadic language serves to scaffold and support children's science exploration (e.g., Vygotsky, 1978).

Our focus on the *delivery* of scientific content is consistent with research indicating that children are not only sensitive to the content of information when making inferences about from whom to learn but are also sensitive to the manner in which that content is delivered (Corriveau and Kurkul, 2014; Mercier et al., 2014; Kurkul and Corriveau, 2018; Mills et al., 2019). Specifically, we explore delivery of scientific content via two different pedagogical approaches they could use when participating in a scientific activity with their child. The first, a *scientific inquiry*–based approach, seeks to leverage children's intuitive drive for seeking explanations through question-asking (Callanan and Oakes, 1992; Kelemen et al., 2014; Frazier et al., 2016; Weisman and Markman, 2017) by guiding them through a process of asking questions, experimenting, and explaining results. Scientific inquiry approaches have been extensively explored and promoted in formal educational contexts (e.g., National Research Council [NRC], 1996, 2001, 2007; Minner et al., 2010; Furtak et al., 2012) and hold promise in informal learning contexts as well (Gutwill and Allen, 2010; Vandermaas-Peeler et al., 2016, 2017, 2018). In this study, we compare this *scientific inquiry*–based approach to a more didactic approach focusing specifically on providing *scientific explanations* to the child without first prompting such explanations through question-asking. Before turning to the current study, we review prior research regarding modification of parents' delivery of scientific content in interactions with their children as well as preliminary studies suggesting benefits in the use of scientific inquiry approaches in informal learning contexts.

Interventions That Modify Parental Delivery of Scientific Language

In the previous section, we highlighted the reasons why parental "business-as-usual" approaches to interacting with their child focus on parents' delivery of explanations that may or may not be scientifically accurate with less of a focus on the scientific process

(e.g., Crowley et al., 2001; Shtulman and Checa, 2012). In recent years, studies of informal science learning have begun to explore other ways parents can deliver scientific content in interactions with their children in museum settings (Fender and Crowley, 2007; Benjamin et al., 2010; Gutwill and Allen, 2010; Haden et al., 2014; Willard et al., 2019). The majority of this research employs "conversation cards" to modify the ways in which parents deliver information to their child. Such conversation cards include printed instructions and prompts and are used as an explicit reminder of how to best discuss information with their child (e.g., Fender and Crowley, 2007; Jant et al., 2014). For example, some interventions include encouraging asking questions (e.g., Benjamin et al., 2010; Haden et al., 2014; Willard et al., 2019), devising or providing explanations (e.g., Fender and Crowley, 2007; Willard et al., 2019), promoting exploration (e.g., Willard et al., 2019), or engaging in multi-step learning processes (e.g., Gutwill and Allen, 2010).

Two recent studies explicitly encouraged parents to increase the number of elaborative *Wh-* questions (what, when, where, why, how) during an engineering-focused museum exhibit (Benjamin et al., 2010; Haden et al., 2014). Both studies found that parents in the question-prompting conditions were more likely to use elaborative *Wh-* questions relative to other conditions. However, *Wh-* question prompting appeared to have mixed impacts on children's own talk. Benjamin et al. (2010) found that children in the questions-prompting instructions conditions were more likely to respond to their parents, to engage in elaborative conversations about engineering concepts, and to correctly recall information about the exhibit. In contrast, Haden et al. (2014) found no such condition differences, suggesting that manipulating parents' delivery in the form of elaborative *Wh-* questions may not always be effective in changing children's own scientific discourse.

A recent study by Willard et al. (2019) focused on employing more minimal conversation-card-based interventions to modify parent-child dyad interactions with a gears activity. Here, the focus was on contrasting (1) exploration of scientific stimuli ("Explore" condition) with (2) explanation of scientific observations ("Explain" condition). Parental questions predicted children's discussion of gears in the Explain condition but not in the Explore condition. To explain this finding, the researchers concluded that, compared to parents in the Explore condition, parents in the Explain condition were likely asking particular types of questions that prompted children to discuss the scientific content at hand. However, as their coding scheme was designed to capture the frequency of questions rather than types of questions (e.g., *Wh-* questions vs. yes/no questions), potential differences in the delivery process across conditions were not explored.

In contrast to using conversation cards to modify specific elements of parental talk, other research has focused on the inquiry process as a whole (Gutwill and Allen, 2010; Vandermaas-Peeler et al., 2016, 2017, 2018). Scientific inquiry (sometimes also referred to as "inquiry science") is most commonly defined as constructivist learning processes wherein children learn from active engagement with scientific activities that focus on observation and experimentation to answer "scientifically-oriented questions" (National Research Council [NRC], 1996, 2001, 2007, 2012; Minner et al., 2010; Furtak et al., 2012; NGSS, 2013). Meta-analyses have indicated that inquiry-based processes have multiple positive effects on children's learning in formal educational contexts, such as increasing engagement in the learning process and drawing conclusions from observations (see Minner et al., 2010; Furtak et al., 2012, for meta-analyses of the efficacy of inquiry-based science teaching).

In informal science learning contexts such as museums, short scientific inquiry-based interventions for family groups have proven effective (Gutwill and Allen, 2010; Vandermaas-Peeler et al., 2016). For example, Vandermaas-Peeler et al. (2016) invited families to interact with museum stations designed to elicit talk about math-centric concepts. As compared to a business-as-usual control group, the families in the group who received informational signs with inquiry-focused suggestions were more likely to produce explaining and reasoning-oriented language. Moreover, children in the inquiry group provided more correct responses to their parents' guidance prompts than the control group, suggesting that children's learning benefitted from participating in inquiry-based informal learning interactions.

Gutwill and Allen (2010) found similarly promising results from a more involved approach to training parents in scientific inquiry processes. Prior to entering the museum exhibit, families participated in "inquiry games" designed to elicit scientific inquiry behaviors. As compared to families in several control conditions, families who participated in the scientific inquiry-based condition, "Juicy Questions," increased the amount of explanations and interpretations when interacting with the exhibits. However, because all analyses were conducted at the level of the family and not separated by parent or child, these findings make it challenging to determine whether the child or the adult was benefitting from the intervention.

Taken together, the results from the studies reviewed above indicate promising but mixed findings for the effectiveness of interventions focusing on scientific delivery (and not just content) in supporting informal science outcomes. Whereas some research (Benjamin et al., 2010; Vandermaas-Peeler et al., 2016; some findings from Willard et al., 2019) indicates benefits to child outcomes, other research (Haden et al., 2014; other conditions from Willard et al., 2019) finds little effects of such delivery. One reason for these mixed results might be associated with limitations of conversation cards as an intervention vehicle. Although conversation cards can prompt specific types of conversational behaviors (e.g., asking more *Wh-* questions), it is challenging to convey effective delivery approaches on a card. Therefore, research on complex interaction approaches such as scientific inquiry may need more involved intervention procedures. To this end, in the current study we chose to use a brief modeled interaction for intervention training, which has been used in previous studies focusing on modifying demonstrated scientific content (Benjamin et al., 2010; Gutwill and Allen, 2010; Haden et al., 2014; Marcus et al., 2017) but has been studied less as a vehicle for modeling delivery processes (Gutwill and Allen, 2010). We explore the impact of this intervention on parental talk and children's talk separately to confirm the effectiveness of such modeling on both dyadic partners. In addition to focusing on

the impact of the intervention on the process through which information is delivered, we also explore impacts on the content of the conversation, as well as potential interactions between delivery and content.

We focused on training parental delivery of scientific *causal* content in an informal museum setting because children's ability to understand underlying mechanisms associated with scientific phenomena is greatly influenced by their ability to explain causal relations (e.g., Gopnik et al., 2004; Legare and Lombrozo, 2014; Walker et al., 2014). Moreover, by the age of three, children begin asking a substantial number of questions focusing on causal inferences (Isaacs, 1930; Callanan and Oakes, 1992; Chouinard et al., 2007; Frazier et al., 2009, 2016; Corriveau and Kurkul, 2014), indicating that children themselves may recognize that causal content is useful for learning about the world around them. In line with this, recently published research has demonstrated that 3- and 4-years-old selectively seek out books that contain "highly causal" as opposed to books that contain "minimally causal" information (Shavlik et al., 2020). Thus, parent talk focusing on highlighting and explaining causal relations between scientific phenomena provides access to information that may be especially engaging to children and support their learning of opaque scientific content.

In the current study, we chose to specifically study children ages 4–6 years of age, as this is an early age range in which children have begun to appreciate the value of causal information (e.g., Chouinard et al., 2007) but also frequently learn in informal learning environments in interactions with their parents (e.g., Willard et al., 2019). Parent-child dyads were randomly assigned to one of two pedagogical training methods. The first, the *scientific inquiry* process approach, invited parents to deliver scientific content via inquiry processes, inspired primarily by the work of Gutwill and Allen's (2010) juicy questions methodology. The second, the *scientific statements* approach, was based on delivering interesting scientific statements in a more didactic manner during a dyadic interaction. This second approach was chosen as didactic approaches are regarded in educational sciences as a more traditional approach to facilitating learning interactions with children and are still being investigated as potential approaches to informal science learning (e.g., Gutwill and Allen, 2010; Willard et al., 2019). By comparing the relative impacts of these approaches on parent-child scientific conversations, we aimed to add to established best practices in assisting parents when they engage in scientific activities with their children. More specifically, this study examined whether training parents in scientific inquiry approaches provides benefits for parent-child conversations beyond the more traditional business-as-usual scientific statements approach.

We asked two research questions. First, we examined the potential impacts of the scientific inquiry and scientific statements interventions on parent-level and child-level scientific talk. Second, we explored the relation between parental talk and children's subsequent talk about the scientific concepts to assess whether this relation differs according to parents' assigned intervention. We predicted that parents who participated in the scientific inquiry approach would be more likely overall to ask questions than parents who observed scientific statement approaches (and conversely, parents who observed the scientific statements approach would use more statements in general). We also predicted that children whose parents observed the scientific inquiry approach would be more likely to verbally respond to their parents' talk and that when they responded, the responses would be more likely to contain scientific content.

MATERIALS AND METHODS

Participants

Fifty-two parents (21 female) and their 4- to 6-years-old children (20 female, M_{age} = 5.43, range = 4.00–6.91) were recruited and tested as dyads in a science museum in the Northeast United States. All participants provided written consent according to standard protocols approved by the institutional review board of the corresponding author's university. An additional five dyads were omitted from the sample; two dyads were omitted because of interference from the children's siblings, two from video file loss, and one from experimenter error. The remaining samples of dyads were randomly assigned to one of two between-subjects conditions, *scientific inquiry* (n = 25) or *scientific statements* (n = 22). On average, parents had a high level of education, consistent with demographics of the average museum visitor ($M_{education}$ = 17.36 years; Soren, 2009; Dawson, 2014). No significant condition differences in education levels were found. Overall, 41 out of 47 parents reported their occupations, of whom 14 (34.15%) had a STEM-related career. No significant differences were found in the percentage of parents who were employed in STEM-related careers between the conditions (*Scientific inquiry*: 38.1%; *Scientific statements*: 30.0%). Average child continuous age (in years) did not significantly differ between conditions, $t(45) = 0.48$, $p = 0.64$.

Testing Procedure

Dyads participated in three phases (baseline, training, post-training). To introduce the tasks, the experimenter said that they were interested in "looking at how children learn through interactions with their caregivers" and asked the dyads to play together as they would at home. Interactions were videotaped for further analysis. Dyads played with three separate activities presented in a fixed order (baseline: a balance scale; training: a circuit board [Snap Circuits© by Elenco]; post-training a mechanical gears layout, see **Figure 1**). The balance scale toy had clear plastic buckets and 74 bear-shaped colored weights. The circuit board (Snap Circuits© by Elenco) consisted of three circuit pieces, a switch, a battery unit, and a lightbulb component arranged in a rectangular circuit on a circuit board. Two additional circuit pieces were on the board but not connected to the circuit. The mechanical gears were based off of Legare and Lombrozo (2014) and consisted of five differently-colored gears of varying sizes (three large and two small) attached to six green hexagonal bases, as well as a crank handle. Recordings of speech samples were transcribed offline. Below, we briefly describe the three tasks and the training conditions before turning to the transcription and coding procedure.

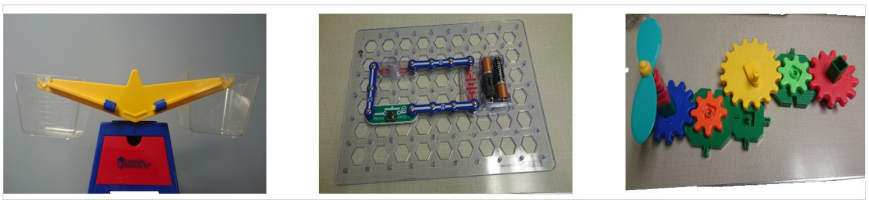


FIGURE 1 | Study phases with their respective tasks.

Baseline

To explore variability in dyadic interaction prior to the intervention, dyads were instructed to "play together, just like you would at home!" with the balance toy. There was no time limit; dyads interacted with the toy as long as they wished before moving on to the next two phases. The average time to complete this phase was 3.84 min (*SD* = 1.70 min, Range = 1.23–8.85 min). Marginally-significant differences were found between the *scientific inquiry* condition (*M* = 3.41 min) and *scientific statements* condition (*M* = 4.34 min) in average interaction time, $t(45) = -1.93$, $p = 0.06$.

Training

Next, the experimenter introduced the dyads to a previously constructed snap circuit board (see **Figure 1**). Dyads watched as the experimenter modeled one of two different pedagogical training conditions using interaction scripts to ensure that all parents received the same training information. As with the baseline, the training phase did not have a time limit; average completion time was 5.44 min (*SD* = 2.87 min, Range = 2.40–17.70 min). Half of the dyads were randomly assigned to the *scientific inquiry* condition. The experimenter introduced the dyad to the snap circuit and its components and explained that they would help with running a scientific investigation. The experimenter described that there were three steps to a "science investigation": question, experiment, and explain. Then, the experimenter explained that in a science investigation, it was important to choose a question that "we don't know the answer to" but could answer with the materials available in the moment. The experimenter invited the child to choose between one of two causal-related questions to guide the inquiry (i.e., "What happens when we take a piece out of the circuit?" or "What happens when we add a piece to the circuit?"). Next, the experimenter prompted the child to make a hypothesis about what would happen and was invited to answer the question through an *experiment* and encouraged to "try it a few times, in different ways" to understand the results. After the child completed their experiment, they were invited to *explain* the answer to the question using the evidence that they had discovered with the experiment.

The remaining half of dyads were presented with the *scientific statements* condition. The experimenter provided the child with information about the different parts of an electrical circuit, taking time to explain specific concepts and mechanisms. For example, the experimenter explained that the word "circuit" was a "fancy word for a loop" and that electricity travels from the battery, around the circuit, and makes the light turn on. The experimenter also demonstrated how pressing the switch connects the circuit, making the light turn on, but releasing the switch disconnects the circuit and makes the light turn off. The child was then informed that it was his or her turn to play and that he or she could use the extra circuit pieces on the board when doing so. Throughout this process, the experimenter explained the results of the child's actions by referencing the previous scientific statements, linking the child's actions and obtained results back to those statements. For example, if a child disconnected a piece from the circuit and then attempted to turn on the light by pressing the switch, the experimenter would explain how all parts of the circuit "loop" had to be connected in order for the light to turn on.

During the administration of both conditions, parents were not actively encouraged or discouraged to participate in the interaction. Regardless of parental participation, the experimenter ensured that all procedures in both conditions were completed so that the parent could observe how to implement the proposed pedagogical approach, whether it was the *scientific inquiry* or *scientific statements* approach.

Post-training

Finally, the experimenter invited the dyad to practice what they had just learned in the training session by playing with the gears activity. The experimenter handed the parent an information sheet for reference, which varied by condition. Both conditions prompted discussion of the same causal content (e.g., the effects of speed and direction on gear movement), but varied according to the delivery of that content. In the *scientific inquiry* condition, parents were prompted to run a science investigation with their child using the "question, experiment, and explain" method, and were provided with four possible questions the dyad might use to explore. In contrast, in the *scientific statements* condition, parents were prompted to share "scientific information" with their children, "just like I did with the circuit," and were provided with a sheet containing scientific explanations about the gears. Parents were explicitly instructed that they were free to use as many or as few questions or statements as they liked when they interacted with their child. As with the previous two phases, there was no time limit; average completion time was 5.04 min (*SD* = 2.79 min, Range = 1.78–17.90 min). Engagement, measured as amount of time spent on the post-training gears task, did not significantly vary by training condition; dyads in the *scientific inquiry* (*M* = 4.65 min) and the *scientific statements* (*M* = 5.49 min) spent about the same amount of time on

the task, $t(45) = -1.04$, $p = 0.31$. Moreover, there were no differences in the amount of child talk. Children in *scientific inquiry* produced 19.76 utterances, and children in *scientific statements* produced 20.73 utterances, on average, $t(45) = -0.17$, $p = 0.87$.

Transcription and Coding

Interactions were transcribed at the level of the utterance for all 47 transcripts. Each transcript was verified by a second research assistant to ensure accuracy. We then removed all utterances that were directly sourced from the text of the supplemental information sheets, yielding a corpus of 3,532 utterances (2,311 parent utterances, 1,221 child utterances). All utterances were examined within the context of the entire transcript, allowing coders to read as much of the interaction as needed to ensure accuracy across the coding categories. All utterances were then coded for their *delivery* and *content*, described in more detail below.

Delivery

The delivery of each parent and child utterance was coded into one of two mutually exclusive categories: *question* (e.g., "How does this work?" P#1, line 21) or *statement* (e.g., "Put this on this one over here." P#18, line 45).

Content

Utterances were initially coded for content, or type of information conveyed by the utterance. Content coding was initially developed based on the causal and fact-based categories used in previous studies (e.g., Chouinard et al., 2007; Frazier et al., 2009; Kurkul and Corriveau, 2018). Overall, eight types of content were coded: fact-based, causal, procedural, confirmation/negation, reinforcement, and irrelevant/unintelligible.

Fact-based

Fact-based utterances included content that discussed scientific facts or observations relevant to the activity. For example, a parent might observe that a gear is spinning fast ("That was super fast." P#13, line 359) or ask how many bears are in one of the balance toy's buckets ("So how many bears was that?" P#20, line 148). These scientific utterances included information from readily observable evidence or from the speaker's prior knowledge of the topic. Note that fact-based statements simply describe an observation about a current state of phenomena and do not provide information or insight into *how* or *why* phenomena or events occur.

Causal

In contrast to fact-based utterances, causal utterances involve discussion of the relation between two or more scientific facts or observations and potentially seek to answer *how* or *why* those phenomena or events occur. These scientific utterances include the relation between cause and effect, a proposed mechanism for change, or a hypothesis or prediction based on proposed actions. For example, a parent might ask a child about why a gear is not spinning with another gear ("Why doesn't it spin with the yellow one?," P#46, line 78) or provide an explanation of why a smaller gear completes more rotations than a big gear when the two are connected ("and if they want to run as fast as you are they have to take a lot more steps' cause they're a lot smaller." P#5, line 133).

Procedural

Utterances were coded as procedural when they provided on-task content regarding a stated goal, steps on what to do next, or other directives that did not contain informational content. These included directives from a parent ("Put the red one there." P#15, line 196) or discussions of intended actions ("Should we try a different one?" P#45, line 54).

Confirmation/negation

These utterances were on-task but low-effort responses that provided little content beyond previous utterances. Examples include "Yes," "No," or "That's right."

Reinforcement

Reinforcement utterances provided motivational feedback or prompting with the goal of continuing the activity or conversation. These include everything from non-confirmatory responses ("okay" as narration of action), positive feedback ("High five, dude." P#13, line 281) to permission granting ("Go ahead.") and conversational fillers ("mm." or "um.").

Irrelevant/unintelligible

Any utterance that was off-task or was uninterpretable on the video recording was coded as either irrelevant or unintelligible.

Proportions of each of these categories by interlocutor (parent, child) and condition are displayed in **Table 1**. Because our main goal was to examine the exchange of substantive content – that is, content that provides relevant information about actions in the scientific activity, observations about the activity, or discussion of scientific phenomena relevant to the activity – the remainder of the analyses focus specifically on *causal, fact-based,* and *procedural* talk.

Reliability

Two research assistants, blind to the training condition and the hypotheses of the study, independently coded the transcripts. Inter-rater reliability was established using 15% of the transcripts. Overall agreement was 84% for content codes (average K = 0.76) and 98% for delivery codes (average K = 0.97). Discrepancies were resolved through discussion.

RESULTS

Based on standard child language data-analytic techniques, we chose to pool data from dyads' talk in the post-training phase, making the utterance, instead of the dyad, the level of the analysis. This approach is consistent with multiple previous studies (e.g., Bartsch et al., 2003; Frazier et al., 2009, 2016; Kurkul and Corriveau, 2018), and is consistent with the standards proposed by Bakeman and Gottman (1997). These standards allow for utterances to be treated as independent, assuming that coding decisions are made separately for individual utterances and the coding categories are mutually exclusive. Both of these conditions were met in the current coding scheme. Moreover, to ensure that

TABLE 1 | Mean percentages of overall talk by content category in the post-training phase by speaker (parent, child) and condition (scientific inquiry, scientific statements).

	Parents			Children		
	Total	**Scientific inquiry**	**Scientific statements**	**Total**	**Scientific inquiry**	**Scientific statements**
Causal %	12.72	14.08	11.18	9.71	11.54	7.62
Fact-based %	21.31	19.15	23.78	24.69	26.42	22.72
Procedural %	29.00	30.71	27.06	18.64	17.43	20.03
Confirmation/Negation %	3.53	3.18	3.94	14.33	11.32	17.75
Reinforcement %	22.46	23.16	21.66	3.30	4.84	1.55
Irrelevant/Unintelligible %	10.88	9.60	12.33	29.33	28.45	30.33

the results were not driven by a few parent-child dyads within a condition, we followed a multi-step analytic process. First, we present mixed-effects models at the level of the utterance to explore each research question. Next, as suggested by Bakeman and Gottman (1997) we confirm that these results hold at the level of the dyad by inspecting the number of participants in each condition that reflect the pattern demonstrated by the mixed-effects models.

To address our first research question, we explore variability in parent-child talk produced in the post-test phase by condition. To examine our second research question, we explore potential relations between parental talk and child-level talk.

Quantity of Talk by Condition

Parents in the *scientific statements* condition (M = 66.41, SD = 37.00) used a marginally larger number of utterances than parents in the *scientific inquiry* condition (M = 49.12, SD = 26.26), $t(45) = 1.86$, $p = 0.07$. Children's total utterances were not significantly different between the *scientific statements* condition (M = 20.73, SD = 20.99) and the *scientific inquiry* condition (M = 19.76, SD = 19.28), $t(45) = 0.17$, $p = 0.87$. Similarly, the amount of overall time dyads spent on task did not differ between the *scientific inquiry* condition (M = 4.65 min, SD = 2.16 min) and the *scientific statements* condition (M = 5.49 min, SD = 3.36 min), $t(45) = -1.04$, $p = 0.31$. To investigate the possibility that the quantity of parent and child talk differed as a function of child age, two linear regressions were run, predicting the total number of child and parent utterances from child age as a continuous variable. The parent model showed a significant effect of child age ($\beta = -12.48$, $SE = 5.38$, $p = 0.03$) on total parent utterances, $F(1, 45) = 5.38$, $p = 0.03$, $R^2 = 0.11$. This indicates that parental talk was negatively associated with child age. The child model did not show a significant effect of child age ($B = -0.18$, $SE = 3.48$, $p = 0.96$) on total child utterances, $F(1, 45) = 0.003$, $p = 0.96$, $R^2 < 0.001$.

Parent Talk by Delivery and Content

To explore the impact of training condition on parents' language, we conducted planned binary-logistic mixed-effects models on the likelihood that a given utterance is a question or statement, with Condition (scientific inquiry, scientific statements; statements was the reference category) and Content (causal, fact-based, procedural; procedural was the reference category for *Model 1* and fact-based was the reference category for *Model 2*) as predictors, and dyad as a random factor to account for individual variance. See **Table 2** (Model 1) for a summary of parent results and **Table 3** (Model 1) for a summary of child results.

Model 1: Effects of Condition by Content

A binary-logistic mixed-effect model on likelihood that a parent's utterance is a question found a significant main effect of

TABLE 2 | Mixed effects binary-logistic regressions on parents' likelihood to ask a question.

				95% CI for *OR*	
Variable	**β (*SE*)**	***Z***	**Odds ratio**	**Lower**	**Upper**
Model 1					
Intercept	0.63 (0.16)***	−3.907	0.54	0.39	0.73
Condition (scientific statements as reference)	0.52 (0.22)*	2.35	1.68	1.09	2.63
Causal Content (procedural as reference)	0.50 (0.20)*	2.50	1.65	1.12	2.45
Fact-based content (procedural as reference)	0.23 (0.16)	1.48	1.26	0.93	1.72
Condition × Causal content	0.90 (0.30)**	3.03	2.46	1.38	4.41
Condition × Fact-based content	−0.23 (0.24)	−0.97	0.80	0.50	1.26
−2LL	−1121.7				
AIC	2257.50				
Model 2					
Intercept	−0.39 (0.16)*	−2.39	0.68	0.49	0.93
Condition (scientific statements as reference)	0.29 (0.24)	1.22	1.34	0.84	2.16
Causal content (fact-based as reference)	0.27 (0.20)	1.33	1.31	0.88	1.95
Procedural content (fact-based as reference)	0.23 (0.24)	−1.48	0.79	0.58	1.08
Condition × Causal content	1.13 (0.31)***	3.64	3.09	1.69	5.70
Condition × Procedural content	0.23 (0.24)	0.97	1.26	0.79	2.00
−2LL	−1121.70				
AIC	2257.50				

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Statements were used as a reference group. CI = 95% confidence interval; OR = odds ratio.*

TABLE 3 | Mixed effects binary logistic regressions on children's likelihood to ask a question.

Variable	β (*SE*)	*Z*	Odds ratio	95% CI for *OR* Lower	Upper
Model 1					
Intercept	−2.16 (0.37)***	−5.85	0.12	0.05	0.22
Condition (scientific statements as reference)	−0.31 (0.57)	−0.54	0.74	0.22	2.15
Causal Content (procedural as reference)	1.83 (0.47)***	3.93	6.26	2.55	16.25
Fact-Based Content (procedural as reference)	−0.34 (0.49)	−0.71	0.71	0.26	1.82
Condition × Causal Content	−1.31 (0.74)[t]	−1.78	0.27	0.06	1.17
Condition × Fact-Based Content	−0.28 (0.74)	−0.38	0.76	0.18	3.29
−2LL	−180.40				
AIC	374.80				
Model 2					
Intercept	−2.50 (0.42)***	−5.976	0.08	0.03	0.17
Condition (scientific statements as reference)	−0.59 (0.59)	−1.00	0.56	0.17	1.77
Causal Content (fact-based as reference)	2.18 (0.53)***	4.09	8.83	3.18	26.36
Procedural Content (fact-based as reference)	0.28 (0.74)	0.71	1.41	0.55	3.80
Condition × Causal Content	−1.03 (0.77)	−1.34	0.36	0.08	1.61
Condition × Procedural Content	0.23 (0.24)	0.70	1.32	0.30	5.56
−2LL	−1121.70				
AIC	2257.50				

[t]*p = 0.08, ***p < 0.001. Statements were used as a reference group. CI, 95% confidence interval; OR, odds ratio.*

TABLE 4 | Estimated likelihood that utterances are a question by condition and delivery.

	Parents		Children	
	Scientific inquiry	Scientific statements	Scientific inquiry	Scientific statements
Causal	0.79	0.47	0.13	0.42
Fact-based	0.48	0.40	0.04	0.08
Procedural	0.47	0.35	0.08	0.10

*Model-estimated mean likelihoods that an utterance of a specific content type is a question (as opposed to a statement) from parent Model 1 (**Table 2**), a binary-logistic mixed-effect model on likelihood that a parent's utterance is a question and child Model 1 (**Table 3**), a binary-logistic mixed-effect model on likelihood that a child's utterance is a question. For instance, in the scientific inquiry condition, the model estimates that there is a 79% likelihood of a causal utterance being posed as a question (as opposed to a statement).*

Condition (β = 0.52, *SE* = 0.22, p = 0.02, *OR* = 1.68, 95% CI [1.09, 2.63]), a significant main effect of Causal Content (B = 0.50, *SE* = 0.20, p = 0.01, *OR* = 1.65, 95% CI [1.12, 2.45]), and a significant interaction between Condition and Causal Content (B = 0.90, *SE* = 0.30, $p < 0.01$, *OR* = 2.46, 95% CI [1.38, 4.41]). No other main effects or interactions were found. Utterances were 1.68 times more likely to be a question if the parent was trained in the *scientific inquiry* condition, 1.65 times more likely to be a question if the parent presented causal content in general, and 2.46 times more likely to be a question if the parent was in the *scientific inquiry* condition and was presenting causal content. See **Table 2** (top panel) for a summary of this model. See **Table 4** for model-estimated mean likelihoods from *Model 1*.

Model 2

To confirm the results of *Model 1*, the reference category for Content was changed to fact-based. This model found a significant interaction between Condition and Causal Content, B = 1.13, *SE* = 0.31, $p < 0.001$, *OR* = 3.09, 95% CI [1.69, 5.70]. Parents in the *scientific inquiry* condition were 3.09 times more likely than parents in the *scientific statements* condition to present causal content as questions. No other main effects or interactions were found. **Table 2** (bottom panel) includes all model parameters.

To confirm that the conclusions of our mixed-model logistic regressions were found across individual dyads, we explored the number of dyads in the *scientific inquiry* condition who displayed more questions than statements when presenting causal content. Twenty-one out of 25 (84%) parents in this condition displayed this pattern. By contrast, only 7 out of 22 (31.81%) parents in the *scientific statements* condition displayed this pattern.

Taken together, both the binary-logistic mixed-models and individual inspection at the dyad level provide similar results. Parents' utterances overall are more likely to be questions when the parent is both in the *scientific inquiry* condition and presenting causal content and a majority of dyads in the *scientific inquiry* condition displayed this pattern.

Child Talk by Delivery and Content

To explore the impact of training condition on children's language, we conducted planned binary-logistic mixed-effects models on the likelihood that a given utterance is a question or statement, with Condition (scientific inquiry, scientific statements; statements was the reference category) and Content (causal, fact-based, procedural; procedural was the reference category for *Model 1* and fact-based was the reference category for *Model 2*) as predictors, and dyad as a random factor to account for individual variance.

Model 1: Effects of Condition by Content

A binary-logistic mixed-effect model on likelihood that a child's utterance is a question found a significant main effect of Causal Content, B = 1.83, *SE* = 0.47, $p < 0.0001$, *OR* = 6.26, 95% CI (2.55, 16.25), and a marginally significant interaction between Condition and Causal Content, β = −1.31, *SE* = 0.74, p = 0.08, *OR* = 0.27, 95% CI [0.06, 1.17]. No other main effects or interactions were found. Compared to procedural talk, children's utterances were 6.26 times more likely to be a question when discussing Causal Content. See **Table 3** (top panel) for a summary

of this model. See **Table 4** for model-estimated mean likelihoods from *Model 1*.

Model 2

To confirm the results of *Model 1*, the reference category for Content was changed to be fact-based. This model found a significant main effect of Causal Content, $B = 2.18$, $SE = 0.53$, $p < 0.0001$, $OR = 8.83$, 95% CI [3.18, 26.36]. Compared to fact-based talk, children's utterances were 8.83 times more likely to be a question when discussing Causal Content. No other main effects or interactions were found. **Table 3** (bottom panel) includes all model parameters.

Child Responses to Parental Causal Utterances

Finally, we explored potential condition-level differences on the relation between the type of parental talk and children's subsequent responses. Recall that parents produced 74% of the overall talk in the post-training phase. We focused specifically on causal talk, as this was the type of talk content that was not only (1) targeted by both conditions but also (2) significantly differed between conditions according to parental delivery. Parental causal utterances were individually coded according to two criteria: (1) whether or not a child provided a "response" (i.e., a child utterance immediately following a parent utterance) and (2) if there was a child response, whether the content of it was scientific in nature (fact-based or causal).

To explore potential condition-level differences, a mixed-model binary logistic regression was run on the likelihood of child responses to parental causal talk with Condition (*scientific inquiry, scientific statements*) as a fixed factor and dyad as a random factor. These results yielded a marginally significant main effect of Condition, $B = 0.84$, $SE = 0.47$, $p = 0.07$, $OR = 2.32$, 95% CI [0.90, 6.42]. No other main effects or interactions were found. Children in the *scientific inquiry* condition were 2.32 times more likely to respond to a parental causal utterance than children in the *scientific statements* condition. **Table 5** includes all model parameters.

To confirm that this pattern held at the level of the dyad, we explored the number of children in each condition who were more likely than not to respond to parents' causal talk (i.e., responded to a parental causal utterance more than 50% of the time). In the *scientific inquiry* condition, 17 out of 25 children in the *scientific inquiry* condition (68.0%) were more likely than not to respond to a parental causal utterance, as compared to 9 out of 21 children (40.9%) in the *scientific statements* condition.

We also explored whether there were Condition differences in the *content* of children's responses to parental causal utterances. Recall that children's responses were coded as "scientific" (causal or fact-based content) or procedural. An additional mixed-model binary logistic regression was run with Condition as a fixed factor and dyad as a random factor. A significant main effect of Condition was found, $\beta = 1.83$, $p = 0.0005$. No other main effects or interactions were found. When responding to parental causal utterances, children in the *scientific inquiry* condition were 6.22 times more likely to provide a scientific response (fact-based or causal) than children in the *scientific statements* condition. **Table 6** includes all model parameters.

To confirm that this pattern held at the level of the dyad, we explored the number of children in each condition who were more likely to respond with a scientific utterance (i.e., provided a scientific response more than 50% of the time) to parent's causal talk. In the *scientific inquiry* condition, 13 out of 25 children in the *scientific inquiry* condition (52%) were more likely than not to provide a scientific response to a parental causal utterance, as compared to 2 out of 21 children (9.52%) in the *scientific statements* condition.

DISCUSSION

In this study, we investigated impacts of a brief pedagogical intervention on both parents' and children's conversation when interacting with scientific activities in a museum context. We also examined how parents' delivery of the causal content targeted by this pedagogical training influenced children's responsiveness to the parent while discussing scientific concepts. Below, we review these findings before discussing implications for parent-child interactions during informal science learning settings, as well as avenues for future research.

Immediately following the short 5 min training, parents' delivery of utterances differed between conditions. However, the content of those utterances did not. When interacting with the gears with their children, parents in both conditions

TABLE 5 | Mixed effects binary logistic regressions on children's likelihood to respond to parental causal talk.

				95% CI for *OR*	
Variable	**β (*SE*)**	***Z***	**Odds ratio**	**Lower**	**Upper**
Intercept	−0.22 (0.35)***	−0.65	0.80	0.37	1.59
Condition (scientific statements as reference)	0.84 (0.47)[t]	1.80	2.32	0.90	6.42
−2LL	−166.90				
AIC	339.90				

[t]$p = 0.08$, ***$p < 0.001$. *No response was used as a reference group. CI, 95% confidence interval; OR, odds ratio.*

TABLE 6 | Mixed effects binary logistic regressions on children's likelihood to respond with scientific talk.

				95% CI for *OR*	
Variable	**β (*SE*)**	***Z***	**Odds ratio**	**Lower**	**Upper**
Intercept	−1.53 (0.42)***	−3.61	0.22	0.08	0.46
Condition (scientific statements as reference)	1.83 (0.53)***	3.47	6.22	2.31	20.80
−2LL	−156.40				
AIC	318.80				

****$p < 0.001$. Procedural content was used as a reference group. Scientific talk was composed of causal or fact-based content. CI, 95% confidence interval; OR, odds ratio.*

engaged in explanatory talk that included causal language, yet the process by which that language was delivered varied by condition. Whereas parents in the *scientific inquiry* condition asked more causal questions, parents in the *scientific statements* condition used more causal statements. Importantly, results from a series of mixed-effects regression models indicated that this condition difference was not found when parents engaged in fact-based or procedural talk. These results indicate that parents' adherence to the modeled pedagogical approaches was specific to the content *targeted* by those approaches – causal scientific concepts, discussing or prompting consideration of the relation between scientific observations, and features of scientific processes.

In addition to impacting parent-level talk, the training session also had effects at the child level. Specifically, children were more likely to respond to parental causal talk in the scientific inquiry condition than in the scientific statements condition. Such responses indicate that parent-child conversation in this condition was more elaborative and included more turns per individual topic. Moreover, when they did respond to their parents' causal talk, children were significantly more likely to produce on-topic scientific responses rather than procedural responses. As a result, parents and children are more likely to generate joint opportunities to learn about specific scientific content in the scientific inquiry condition. This finding complements and extends the research by Gutwill and Allen (2010) and Vandermaas-Peeler et al. (2016) by highlighting how inquiry processes can prompt extended conversations about science between parents and children, illustrating the important role that social interaction plays for children's scientific learning. Indeed, recent research has highlighted the importance of extended conversational turns with parents for children's language outcomes (Romeo et al., 2018). Additionally, this study builds on previous work finding that children are more likely to respond to *Wh-* questions, which include causal questions, as compared to other types of questions (Rowe et al., 2017). The current study highlights how conversational turns might also be important for strengthening children's domain knowledge. However, it is important to note that the low child response rate overall prevented further exploration via mediation analyses into the extent to which aspects of both the delivery and content most affected child response rates. Future research should explore this question further.

Parents' willingness to adopt both the delivery and content targeted by the short modeled training is consistent with previous training studies focusing on modifying specific targeted language (e.g., *Wh-* questions: Benjamin et al., 2010; Haden et al., 2014) or modifying inquiry processes at the family level (Gutwill and Allen, 2010). As we indicate above, our findings replicate and extend this work to highlight that such guidance has effects at both the parent level and the child level and not only for specific content but also for the process through which that content is delivered. Importantly, findings from the current study were found in a sample of parent-child dyads with children who were considerably younger than those included in previous informal science learning interventions (e.g., Gutwill and Allen, 2010; Vandermaas-Peeler et al., 2016, 2017, 2018). Thus, inquiry-based approaches appear to be effective during the developmental period when parents are children's primary teachers – prior to their child's extensive experiences with formal education.

The impact of the short training session on parent-level and child-level talk is all the more striking when considering that the training generalized across tasks (from an electrical circuit to gears) and was found in spontaneous, self-generated speech. Recall that any utterances that specifically referenced the post-training supplemental information sheet were not included in analyses. These data are consistent with some recent research indicating impacts of parental causal talk on children's ability to generalize causal mechanisms when interacting with novel physical stimuli (Kurkul et al., under review; Leech et al., 2019). Future research should explore how such an intervention not only impacts the explanatory talk by parents and children but also impacts child-level learning outcomes.

This study has several notable limitations. The first was that the effectiveness of the short training on parent-child conversation was explored immediately following the training session. Some research has indicated impacts of interventions up to 1 month later (Gutwill and Allen, 2010; Haden, 2010; Haden et al., 2014) and so it is plausible that such short interventions can have longer-lasting effects. Future research should explore the extent to which dyads continue to show training effectiveness on parent-child conversation with a time delay, or across other contexts beyond the museum. Second, as we tested our intervention in the context of a gears activity, it is unknown whether our scientific inquiry approach would be more or less successful across other science activities. Third, as this study aimed to compare the efficacy of the scientific inquiry and scientific statements approaches in eliciting child engagement in scientific conversations, it did not utilize a pre-test–post-test design or a direct measure of child learning outcomes. As such, our findings do not reveal whether parents' natural interaction styles were modified by the intervention or whether children learned specific scientific concepts or facts.

Finally, because data were collected from a sample of parent-child dyads in a science museum, most participants likely had an interest in science and the majority were likely to be of middle-to-high socioeconomic status. Future research should explore the effectiveness of these training interventions on a more diverse sample of families. Dyadic inquiry processes and conversations vary considerably based on family background (Tizard and Hughes, 1984; Bang and Medin, 2010; Coppens et al., 2014; Rogoff, 2014; Solis and Callanan, 2016; Kurkul and Corriveau, 2018; Gauvain and Munroe, 2020). Therefore, the inquiry processes modeled here may be less familiar to some families, making the practices more or less difficult to integrate into typical conversational patterns. Indeed, family inquiry can take many forms, and may not always be displayed in the manner we have defined inquiry in this study. In some communities, a different type of inquiry might be displayed as children "listening in" or engaging in non-verbal observation of others (e.g., Rogoff, 2003). In these or other communities, question-asking to adults is not common or expected (Gauvain, 2001; Johnston and Wong, 2002;

Rogoff, 2003; Gauvain and Munroe, 2020). Future research should explore how best to adapt this training session to the discourse patterns of diverse family backgrounds to promote engaging scientific conversations between parents and children.

Taken together, this study provides evidence that a brief training in scientific inquiry helps parents and children talk about higher-order scientific causal content in a collaborative manner in the context of a science museum. Training parents to talk with their children regarding the causal relations underpinning scientific concepts has the potential to have far-reaching impacts on children's interest in and learning of scientific concepts prior to their introduction to formal schooling. Future research should further explore the full implications of scientific inquiry training and other similar methodologies on parent-child dyadic talk in informal learning settings.

DATA AVAILABILITY STATEMENT

A de-identified version of the dataset used for multi-level analyses is available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Boston University Institutional Review Board, Charles River Campus. Written informed consent to participate in this study was provided by the participants' legal guardian.

AUTHOR CONTRIBUTIONS

IC-C wrote the first draft of the manuscript and collected the data. KL and KC provided substantial feedback. KC and IC-C designed the study. IC-C, KL, and KC designed the coding and transcription schemes, and conducted data analysis. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the grant from the National Science Foundation (EHR 1652224) to KC.

REFERENCES

Bakeman, R., and Gottman, J. M. (1997). *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge University Press.

Bang, M., and Medin, D. (2010). Cultural processes in science education: supporting the navigation of multiple epistemologies. *Sci. Educ.* 94, 1008–1026. doi: 10.1002/sce.20392

Bartsch, K., Horvath, K., and Estes, D. (2003). Young children's talk about learning events. *Cogn. Dev.* 18, 177–193. doi: 10.1016/S0885-2014(03)00019-4

Benjamin, N., Haden, C. A., and Wilkerson, E. (2010). Enhancing building, conversation, and learning through caregiver–child interactions in a children's museum. *Dev. Psychol.* 46, 502–515. doi: 10.1037/a0017822

Callanan, M. A., and Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: causal thinking in everyday activity. *Cogn. Dev.* 7, 213–233. doi: 10.1016/0885-2014(92)90012-g

Chouinard, M. M., Harris, P. L., and Maratsos, M. P. (2007). Children's questions: a mechanism for cognitive development. *Monogr. Soc. Res. Child Dev.* 72, vii–ix,1–112; discussion113–26.

Coley, J. D., and Tanner, K. (2015). Relations between intuitive biological thinking and biological misconceptions in biology majors and nonmajors. *CBE Life Sci. Educ.* 14:ar8. doi: 10.1187/cbe.14-06-0094

Coley, J. D., and Tanner, K. D. (2012). Common origins of diverse misconceptions: cognitive principles and the development of biology thinking. *CBE Life Sci. Educ.* 11, 209–215. doi: 10.1187/cbe.12-06-0074

Coppens, A., Silva, K., Ruvalcaba, O., Alcalá, L., López, A., and Rogoff, B. (2014). Learning by observing and pitching in: benefits and processes of expanding repertoires. *Hum. Dev.* 57, 150–161. doi: 10.1159/000356770

Corriveau, K. H., and Kurkul, K. E. (2014). "Why does rain fall?" Children prefer to learn from an informant who uses noncircular explanations. *Child Dev.* 85, 1827–1835.

Crowley, K., and Callanan, M. (1998). Describing and supporting collaborative scientific thinking in parent-child interactions. *J. Museum Educ.* 23, 12–17. doi: 10.1080/10598650.1998.11510365

Crowley, K., Callanan, M. A., Jipson, J. L., Galco, J., Topping, K., and Shrager, J. (2001). Shared scientific thinking in everyday parent-child activity. *Sci. Educ.* 85, 712–732. doi: 10.1002/sce.1035

Dawson, E. (2014). "Not designed for us": How science museums and science centers socially exclude low-income, minority ethnic groups. *Sci. Educ.* 98, 981–1008. doi: 10.1002/sce.21133

Dotterer, A. M., McHale, S. M., and Crouter, A. C. (2009). The development and correlates of academic interests from childhood through adolescence. *J. Educ. Psychol.* 101, 509–519. doi: 10.1037/a0013987

Fender, J. G., and Crowley, K. (2007). How parent explanation changes what children learn from everyday scientific thinking. *J. Appl. Dev. Psychol.* 28, 189–210. doi: 10.1016/j.appdev.2007.02.007

Frazier, B. N., Gelman, S. A., and Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult-child conversation. *Child Dev.* 80, 1592–1161.

Frazier, B. N., Gelman, S. A., and Wellman, H. M. (2016). Young children prefer and remember satisfying explanations. *J. Cogn. Dev.* 17, 718–736. doi: 10.1080/15248372.2015.1098649

Furtak, E. M., Seidel, T., Iverson, H., and Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: a meta-analysis. *Rev. Educ. Res.* 82, 300–329. doi: 10.3102/0034654312457206

Gauvain, M. (2001). *The Social Context of Cognitive Development*. New York, NY: Guilford Press.

Gauvain, M., and Munroe, R. L. (2020). "10 children's questions in social and cultural perspective," in *The Questioning Child: Insights from Psychology and Education*, 183, eds L. P. Butler, S. Ronfard, and K. H. Corriveau (Cambridge: Cambridge University Press).

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychol. Rev.* 111, 3–32. doi: 10.1037/0033-295x.111.1.3

Gutwill, J. P., and Allen, S. (2010). Facilitating family group inquiry at science museum exhibits. *Sci. Educ.* 94, 710–742. doi: 10.1002/sce.20387

Haden, C. A. (2010). Talking about science in museums. *Child Dev. Perspect.* 4, 62–67. doi: 10.1111/j.1750-8606.2009.00119.x

Haden, C. A., Jant, E. A., Hoffman, P. C., Marcus, M., Geddes, J. R., and Gaskins, S. (2014). Supporting family conversations and children's STEM learning in a children's museum. *Early Child. Res. Q.* 29, 333–344.

Isaacs, N. (1930). "Children's "why" questions," in *Intellectual Growth in Young Children*, ed. S. Isaacs (London: George Routledge & Sons).

Jacobs, J. E., and Bleeker, M. M. (2004). Girls' and boys' developing interests in math and science: do parents matter? *New Direct. Child Adoles. Dev.* 2004, 5–21. doi: 10.1002/cd.113

Jant, E. A., Haden, C. A., Uttal, D. H., and Babcock, E. (2014). Conversation and object manipulation influence children's learning in a museum. *Child Dev.* 85, 2029–2045. doi: 10.1111/cdev.12252

Jipson, J. L., and Callanan, M. A. (2003). Mother–child conversation and children's understanding of biological and nonbiological changes in size. *Child Dev.* 74, 629–644. doi: 10.1111/1467-8624.7402020

Johnston, J. R., and Wong, M. Y. A. (2002). Cultural differences in beliefs and practices concerning talk to children. *J. Speech Lang. Hear. Res.* 45, 916–926. doi: 10.1044/1092-4388(2002/074)

Kelemen, D., Emmons, N. A., Seston Schillaci, R., and Ganea, P. A. (2014). Young children can be taught basic natural selection using a picture-storybook intervention. *Psychol. Sci.* 25, 893–902. doi: 10.1177/0956797613516009

Kelemen, D., and Rosset, E. (2009). The human function compunction: teleological explanation in adults. *Cognition* 111, 138–143. doi: 10.1016/j.cognition.2009.01.001

Kurkul, K. E., and Corriveau, K. H. (2018). Question, explanation, follow-up: a mechanism for learning from others? *Child Dev.* 89, 280–294. doi: 10.1111/cdev.12726

Leech, K. A., Campbell, I. L., Alton, J., and Corriveau, K. H. (2019). "What would happen if?: A comparison of fathers' and mothers' questions to children in the museum," *Paper Presented at the Biennial Society for Research in Child Development*, Baltimore, MD.

Legare, C. H., and Lombrozo, T. (2014). Selective effects of explanation on early learning during childhood. *J. Exp. Child Psychol.* 126, 198–212. doi: 10.1016/j.jecp.2014.03.001

Legare, C. H., Sobel, D. M., and Callanan, M. (2017). Causal learning is collaborative: examining explanation and exploration in social contexts. *Psychon. Bull. Rev.* 24, 1548–1554. doi: 10.3758/s13423-017-1351-3

Leichtman, M. D., Camilleri, K. A., Pillemer, D. B., Amato-Wierda, C. C., Hogan, J. E., and Dongo, M. D. (2017). Talking after school: parents' conversational styles and children's memory for a science lesson. *J. Exp. Child Psychol.* 156, 1–15. doi: 10.1016/j.jecp.2016.11.002

Marcus, M., Haden, C. A., and Uttal, D. H. (2017). STEM learning and transfer in a children's museum and beyond. *Merrill Palmer Q.* 63, 155–180.

Mercier, H., Bernard, S., and Clément, F. (2014). Early sensitivity to arguments: how preschoolers weight circular arguments. *J. Exp. Child Psychol.* 125, 102–109. doi: 10.1016/j.jecp.2013.11.011

Mills, C. M., Sands, K. R., Rowles, S. P., and Campbell, I. L. (2019). "I Want to Know More!": children are sensitive to explanation quality when exploring new information. *Cogn. Sci.* 43:e12706. doi: 10.1111/cogs.12706

Minner, D. D., Levy, A. J., and Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *J. Res. Sci. Teach.* 47, 474–496. doi: 10.1002/tea.20347

National Research Council [NRC] (1996). *National Science Education Standards*. Washington, DC: National Academies Press.

National Research Council [NRC] (2001). *Inquiry and the National Science Education Standards*. Washington, DC: National Academies Press.

National Research Council [NRC] (2007). *Taking Science to School: Learning and Teaching Science in Grades K–8*. Washington, DC: National Academies Press.

National Research Council [NRC] (2012). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: The National Academies Press.

NGSS, (2013). *The Next Generation Science Standards: Executive Summary*. Washington, DC: National Center for Education Statistics.

Rigney, J. C., and Callanan, M. A. (2011). Patterns in parent–child conversations about animals at a marine science center. *Cogn. Dev.* 26, 155–171. doi: 10.1016/j.cogdev.2010.12.002

Rogoff, B. (2003). *The Cultural Nature of Human Development*. Oxford: Oxford University Press.

Rogoff, B. (2014). Learning by observing and pitching in to family and community endeavors: an orientation. *Hum. Dev.* 57, 69–81. doi: 10.1159/000356757

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., et al. (2018). Beyond the 30-million-word gap: Children's conversational exposure is associated with language-related brain function. *Psychol. Sci.* 29, 700–710. doi: 10.1177/0956797617742725

Rowe, M. L., Leech, K. A., and Cabrera, N. (2017). Going beyond input quantity: Wh-questions matter for toddlers' language and cognitive development. *Cogn. Sci.* 41, 162–179. doi: 10.1111/cogs.12349

Shavlik, M., Bauer, J. R., and Booth, A. E. (2020). Children's preference for causal information in storybooks. *Front. Psychol.* 11:166. doi: 10.3389/fpsyg.2020.00666

Shtulman, A., and Checa, I. (2012). Parent-child conversations about evolution in the context of an interactive museum display. *Int. Electron. J. Element. Educ.* 5:27.

Solis, G., and Callanan, M. (2016). Evidence against deficit accounts: conversations about science in Mexican heritage families living in the United States. *Mind Cult. Activ.* 23, 212–224. doi: 10.1080/10749039.2016.1196493

Soren, B. J. (2009). Museum experiences that change visitors. *Museum Management and Curatorship* 24, 233–251. doi: 10.1080/09647770903073060

Tenenbaum, H. R., Snow, C. E., Roach, K. A., and Kurland, B. (2005). Talking and reading science: longitudinal data on sex differences in mother–child conversations in low-income families. *J. Appl. Dev. Psychol.* 26, 1–19. doi: 10.1016/j.appdev.2004.10.004

Tizard, B., and Hughes, M. (1984). *Children Learning at Home and in School*. London: Fontana.

Vandermaas-Peeler, M., Massey, K., and Kendall, A. (2016). Parent guidance of young children's scientific and mathematical reasoning in a science museum. *Early Child. Educ. J.* 44, 217–224. doi: 10.1007/s10643-015-0714-5

Vandermaas-Peeler, M., Mischka, M., and Sands, K. (2017). 'What do you notice?'Parent guidance of preschoolers' inquiry in activities at home. *Early Child Dev. Care* 189, 220–232. doi: 10.1080/03004430.2017.1310724

Vandermaas-Peeler, M., Westerberg, L., Fleishman, H., Sands, K., and Mischka, M. (2018). Parental guidance of young children's mathematics and scientific inquiry in games, cooking, and nature activities. *Int. J. Early Years Educ.* 5, 27–46.

Vlach, H. A., and Noll, N. (2016). Talking to children about science is harder than we think: characteristics and metacognitive judgments of explanations provided to children and adults. *Metacogn. Learn.* 11, 317–338. doi: 10.1007/s11409-016-9153-y

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Walker, C. M., Lombrozo, T., Legare, C. H., and Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition* 133, 343–357. doi: 10.1016/j.cognition.2014.07.008

Wang, M. T., and Degol, J. (2013). Motivational pathways to STEM career choices: using expectancy–value perspective to understand individual and gender differences in STEM fields. *Dev. Rev.* 33, 304–340. doi: 10.1016/j.dr.2013.08.001

Weisman, K., and Markman, E. M. (2017). Theory-based explanation as intervention. *Psychon. Bull. Rev.* 24, 1555–1562. doi: 10.3758/s13423-016-1207-2

Willard, A. K., Busch, J. T. A., Cullum, K. A., Letourneau, S. M., Sobel, D. M., Callanan, M., et al. (2019). Explain this, explore that: a study of parent–child interaction in a children's museum. *Child Dev.* 90, e598–e617.

Yu, Y., Bonawitz, E., and Shafto, P. (2019). Pedagogical questions in parent-child conversations. *Child Dev.* 90, 147–161. doi: 10.1111/cdev.12850

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Chandler-Campbell, Leech and Corriveau. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

PERSPECTIVE
published: 05 August 2020
doi: 10.3389/fpsyg.2020.01717

Supporting Early Scientific Thinking Through Curiosity

*Jamie J. Jirout**

Curry School of Education and Human Development, University of Virginia, Charlottesville, VA, United States

Curiosity and curiosity-driven questioning are important for developing scientific thinking and more general interest and motivation to pursue scientific questions. Curiosity has been operationalized as preference for uncertainty (Jirout and Klahr, 2012), and engaging in inquiry-an essential part of scientific reasoning-generates high levels of uncertainty (Metz, 2004; van Schijndel et al., 2018). This perspective piece begins by discussing mechanisms through which curiosity can support learning and motivation in science, including motivating information-seeking behaviors, gathering information in response to curiosity, and promoting deeper understanding through connection-making related to addressing information gaps. In the second part of the article, a recent theory of how to promote curiosity in schools is discussed in relation to early childhood science reasoning. Finally, potential directions for research on the development of curiosity and curiosity-driven inquiry in young children are discussed. Although quite a bit is known about the development of children's question asking specifically, and there are convincing arguments for developing scientific curiosity to promote science reasoning skills, there are many important areas for future research to address how to effectively use curiosity to support science learning.

Keywords: curiosity, scientific reasoning, scientific thinking, information seeking, exploration, learning

OPEN ACCESS

Edited by:
Ageliki Nicolopoulou, Lehigh University, United States

Reviewed by:
Pierre-Yves Oudeyer, Institut National de Recherche en Informatique et en Automatique (INRIA), France
Patricia A. Ganea, University of Toronto, Canada

***Correspondence:**
Jamie J. Jirout
jirout@virginia.edu

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *28 February 2020*
Accepted: *23 June 2020*
Published: *05 August 2020*

Citation:
Jirout JJ (2020) Supporting Early Scientific Thinking Through Curiosity. Front. Psychol. 11:1717. doi: 10.3389/fpsyg.2020.01717

SCIENTIFIC THINKING AND CURIOSITY

Scientific thinking is a type of knowledge seeking involving intentional information seeking, including asking questions, testing hypotheses, making observations, recognizing patterns, and making inferences (Kuhn, 2002; Morris et al., 2012). Much research indicates that children engage in this information-seeking process very early on through questioning behaviors and exploration. In fact, children are quite capable and effective in gathering needed information through their questions, and can reason about the effectiveness of questions, use probabilistic information to guide their questioning, and evaluate who they should question to get information, among other related skills (see Ronfard et al., 2018 for review). Although formal educational contexts typically give students questions to explore or steps to follow to "do science," young children's scientific thinking is driven by natural curiosity about the world around them, and the desire to understand it and generate their own questions about the world (Chouinard et al., 2007; Duschl et al., 2007; French et al., 2013; Jirout and Zimmerman, 2015).

What Does Scientific Curiosity Look Like?

Curiosity is defined here as the desire to seek information to address knowledge gaps resulting from uncertainty or ambiguity (Loewenstein, 1994; Jirout and Klahr, 2012). Curiosity is often seen

as ubiquitous within early childhood. Simply observing children can provide numerous examples of the bidirectional link between curiosity and scientific reasoning, such as when curiosity about a phenomenon leads to experimentation, which, in turn, generates new questions and new curiosities. For example, an infant drops a toy to observe what will happen. When an adult stoops to pick it up, the infant becomes curious about how many times an adult will hand it back before losing interest. Or, a child might observe a butterfly over a period of time, and wonder why it had its wings folded or open at different points, how butterflies fly, why different butterflies are different colors, and so on (see **Figure 1**). Observations lead to theories, which may be immature, incomplete, or even inaccurate, but so are many early scientific theories. Importantly, theories can help identify knowledge gaps, leading to new instances of curiosity and motivating children's information seeking to acquire new knowledge and, gradually, correct misconceptions. Like adults, children learn from their experiences and observations and use information about the probability of events to revise their theories (Gopnik, 2012).

Although this type of reasoning is especially salient in science, curiosity can manifest in many different types of information seeking in response to uncertainty, and is similar to critical thinking in other domains of knowledge and to active learning and problem solving more generally (Gopnik, 2012; Klahr et al., 2013; Saylor and Ganea, 2018). The development of scientific thinking begins as the senses develop and begin providing information about the world (Inhelder and Piaget, 1958; Gopnik et al., 1999). When they are not actively discouraged, children need no instruction to ask questions and explore, and the information they get often leads to further information seeking. In fact, observational research suggests that children can ask questions at the rate of more than 100 per hour (Chouinard et al., 2007)! Although the adults in a child's life might tire of what seems like relentless questioning (Turgeon, 2015), even young children can modify their beliefs and learn from the information they receive (Ronfard et al., 2018). More generally, children seek to understand their world through active exploration, especially in response to recognizing a gap in their understanding (Schulz and Bonawitz, 2007). The active choice of what to learn, driven by curiosity, can provide motivation and meaning to information and instill a lasting positive approach to learning in formal educational contexts.

FIGURE 1 | A child looks intently at a butterfly, becoming curious about the many things she wonders based on her observations.

How Does Curiosity Develop and Support Scientific Thinking?

There are several mechanisms through which children's curiosity can support the development and persistence of scientific thinking. Three of these are discussed below, in sequence: that curiosity can (1) motivate information-seeking behavior, which leads to (2) question-asking and other information-seeking behaviors, which can (3) activate related previous knowledge and support deeper learning. Although we discuss these as independent, consecutive steps for the sake of clarity, it is much more likely that curiosity, question asking and information seeking, and cognitive processing of information and learning are all interrelated processes that support each other (Oudeyer et al., 2016). For example, information seeking that is not a result of curiosity can lead to new questions, and as previous knowledge is activated it may influence the ways in which a child seeks information.

Curiosity as a Motivation for Information Seeking

Young children's learning is driven by exploration to make sense of the world around them (e.g., Piaget, 1926). This exploration can result from curiosity (Loewenstein, 1994; Jirout and Klahr, 2012) and lead to active engagement in learning (Saylor and Ganea, 2018). In the example given previously, the child sees that some butterflies have open wings and some have closed wings, and may be uncertain about why, leading to more careful observations that provide potential for learning. Several studies demonstrate that the presence of uncertainty or ambiguity leads to higher engagement (Howard-Jones and Demetriou, 2009) and more exploration and information seeking (Berlyne, 1954; Lowry and Johnson, 1981; Loewenstein, 1994; Litman et al., 2005; Jirout and Klahr, 2012). For example, when children are shown ambiguous demonstrations for how a novel toy works, they prefer and play longer with that toy than with a new toy that was demonstrated without ambiguity (Schulz and Bonawitz, 2007). Similar to ambiguity, surprising or unexpected observations can create uncertainty and lead to curiosity-driven questions or explanations through adult–child conversations (Frazier et al., 2009; Danovitch and Mills, 2018; Jipson et al., 2018). This curiosity can promote lasting effects; Shah et al. (2018) show that young children's curiosity, reported by parents at the start of kindergarten, relates to academic school readiness. In one of the few longitudinal studies including curiosity, research shows that parents' promotion of curiosity early in childhood leads to science intrinsic motivation years later and science achievement in high school (Gottfried et al., 2016). More generally, curiosity can provide a remedy to boredom, giving children a goal to direct their behavior and the motivation to act on their curiosity (Litman and Silvia, 2006).

Curiosity as Support for Directing Information-Seeking Behavior

Gopnik et al. (2015) suggest that adults are efficient in their attention allocation, developed through extensive experience, but this attentional control comes at the cost of missing much of what is going on around them unrelated to their goals. Children have less experience and skill in focusing their attention, and more exploration-oriented goals, resulting in more open-ended exploratory behavior but also more distraction. Curiosity can help focus children's attention on the specific information being sought (e.g., Legare, 2014). For example, when 7–9-year-old children completed a discovery-learning task in a museum, curiosity was related to more efficient learning-more curious children were quicker and learned more from similar exploration than less-curious children (van Schijndel et al., 2018). Although children are quite capable of using questions to express curiosity and request specific information (Berlyne, 1954; Chin and Osborne, 2010; Jirout and Zimmerman, 2015; Kidd and Hayden, 2015; Luce and Hsi, 2015), these skills can and should be strategically supported, as question asking plays a fundamental role in science and is important to develop (Chouinard et al., 2007; Dewey, 1910; National Governors Association, 2010; American Association for the Advancement of Science [AAAS], 1993; among others). Indeed, the National Resource Council (2012) National Science Education Standards include question asking as the *first* of eight scientific and engineering practices that span all grade levels and content areas.

Children are proficient in requesting information from quite early ages (Ronfard et al., 2018). Yet, there are limitations to children's question asking; it can be "inefficient." For example, to identify a target object from an array, young children often ask confirmation questions or make guesses rather than using more efficient "constraint-seeking" questions (Mills et al., 2010; Ruggeri and Lombrozo, 2015). However, this behavior is observed in highly structured problem-solving tasks, during which children likely are not very curious. In fact, if the environment contains other things that children are curious about, it could be more efficient to use a simplistic strategy, freeing up cognitive resources for the true target of their curiosity. More research is needed to better understand children's use of curiosity-driven questioning behavior as well as exploration, but naturalistic observations show that children do ask questions spontaneously to gain information, and that their questions (and follow-up questions) are effective in obtaining desired information (Nelson et al., 2004; Kelemen et al., 2005; Chouinard et al., 2007).

Curiosity as Support for Deeper Learning

Returning to the definition of curiosity as information seeking to address knowledge gaps, becoming curious-by definition-involves the activation of previous knowledge, which enhances learning (VanLehn et al., 1992; Conati and Carenini, 2001). The active learning that results from curiosity-driven information seeking involves meaningful cognitive engagement and constructive processing that can support deeper learning (Bonwell and Eison, 1991; King, 1994; Loyens and Gijbels, 2008). The constructive process of seeking information to generate new thinking or new knowledge in response to curiosity is a more effective means of learning than simply receiving information (Chi and Wylie, 2014). Even if information is simply given to a child as a result of their asking a question, the mere process of recognizing the gap in one's knowledge to have a question activates relevant previous knowledge and leads to more effective storage of the new information within a meaningful mental representation; the generation of the question is a constructive process in itself. Further, learning more about a topic allows children to better recognize their related knowledge and information gaps (Danovitch et al., 2019). This metacognitive reasoning supports learning through the processes of activating, integrating, and inferring involved in the constructive nature of curiosity-drive information seeking (Chi and Wylie, 2014). Consistent with this theory, Lamnina and Chase (2019) showed that higher curiosity, which increased with the amount of uncertainty in a task, related to greater transfer of middle school students' learning about specific science topics.

PROMOTING CURIOSITY IN YOUNG CHILDREN

Curiosity is rated by early childhood educators as "very important" or "essential" for school readiness and considered to be even more important than discrete academic skills like counting and knowing the alphabet (Heaviside et al., 1993; West et al., 1993), behind only physical health and communication skills in importance (Harradine and Clifford, 1996). Engel (2011, 2013) finds that curiosity declines with development and suggests that understanding how to promote or at least sustain it is important. Although children's curiosity is considered a natural characteristic that is present at birth, interactions with and responses from others can likely influence curiosity, both at a specific moment and context and as a more stable disposition (Jirout et al., 2018). For example, previous work suggests that curiosity can be promoted by encouraging children to feel comfortable with and explore uncertainty (Jirout et al., 2018); experiences that create uncertainty lead to higher levels of curious behavior (e.g., Bonawitz et al., 2011; Engel and Labella, 2011; Gordon et al., 2015).

One strategy for promoting curiosity is through classroom climate; children should feel safe and be encouraged to be curious and exploration and questions should be valued (Pianta et al., 2008). This is accomplished by de-emphasizing being "right" or all-knowing, and instead embracing uncertainty and gaps in one's own knowledge as opportunities to learn. Another strategy to promote curiosity is to provide support for the information-seeking behaviors that children use to act on their curiosity. There are several specific strategies that may promote children's curiosity (see Jirout et al., 2018, for additional strategies), including:

1. Encourage and provide opportunities for children to explore and "figure out," emphasizing the value of the process (exploration) over the outcome (new knowledge or skills). Children cannot explore if opportunities are

not provided to them, and they will not ask questions if they do not feel that their questions are welcomed. Even if opportunities and encouragement are provided, the fear of being wrong can keep children from trying to learn new things (Martin and Marsh, 2003; Martin, 2011). Active efforts to discover or "figure out" are more effective at supporting learning than simply telling children something or having them practice learned procedures (Schwartz and Martin, 2004). Children can explore when they have guidance and support to engage in think-aloud problem solving, instead of being told what to try or getting questions answered directly (Chi et al., 1994).

2. Model curiosity for children, allowing them to see that others have things that they do not know and want to learn about, and that others also enjoy information-seeking activities like asking questions and researching information. Technology makes information seeking easier than it has ever been. For example, children are growing up surrounded by internet-connected devices (more than 8 per capita in 2018), and asking questions is reported to be one of the most frequent uses of smart speakers (NPR-Edison Research Spring, 2019). Observing others seeking information as a normal routine can encourage children's own question asking (McDonald, 1992).
3. Children spontaneously ask questions, but adults can encourage deeper questioning by using explicit prompts and then supporting children to generate questions (King, 1994; Rosenshine et al., 1996). This is different from asking "Do you have any questions?," which may elicit a simple "yes" or "no" response from the child. Instead, asking, "What questions do you have?" is more likely to provide a cue for children to practice analyzing what they do not know and generating questions. The ability to evaluate one's knowledge develops through practice, and scaffolding this process by helping children recognize questions to ask can effectively support development (Kuhn and Pearsall, 2000; Chin and Brown, 2002).
4. Other methods to encourage curiosity include promoting and reinforcing children's thinking about alternative ideas, which could also support creativity. Part of being curious is recognizing questions that can be asked, and if children understand that there are often multiple solutions or ways to do something they will be more likely to explore to learn "*how* we know and *why* we believe; e.g., to expose science as a way of knowing" (Duschl and Osborne, 2002, p. 40). Children who learn to "think outside the box" will question what they and others know and better understand the dynamic nature of knowledge, supporting a curious mindset (Duschl and Osborne, 2002).

Although positive interactions can promote and sustain curiosity in young children, curiosity can also be suppressed or discouraged through interactions that emphasize performance or a focus on explicit instruction (Martin and Marsh, 2003; Martin, 2011; Hulme et al., 2013). Performance goals, which are goals that are focused on demonstrating the attainment of a skill, can lead to lower curiosity to avoid distraction or risk to achieving the goal (Hulme et al., 2013). Mastery goals, which focus on understanding and the learning process, support learning for its own sake (Ames, 1993). When children are older and attend school, they experience expectations that prioritize performance metrics over academic and intellectual exploration, such as through tests and state-standardized assessments, which discourages curiosity (Engel, 2011; Jirout et al., 2018). In my own recent research, we observed a positive association between teachers' use of mastery-focused language and their use of curiosity-promoting instructional practices in preschool math and science lessons (Jirout and Vitiello, 2019). Among 5th graders, student ratings of teacher emphasis on standardized testing was associated with lower observed curiosity-promotion by teachers (Jirout and Vitiello, 2019). It is likely that learning orientations influence children's curiosity even before children begin formal schooling, and de-emphasizing performance is a way to support curiosity.

In summary, focusing on the process of "figuring out" something children do not know, modeling and explicitly prompting exploration and question asking, and supporting metacognitive and creative thinking are all ways to promote curiosity and support effective cognitive engagement during learning. These methods are consistent with inquiry-based and active learning, which both are grounded in constructivism and information gaps similar to the current operationalization of curiosity (Jirout and Klahr, 2012; Saylor and Ganea, 2018; van Schijndel et al., 2018). Emphasizing performance, such as academic climates focused on teaching rote procedures and doing things the "correct" way to get the right answer, can suppress or discourage curiosity. Instead, creating a supportive learning climate and responding positively to curiosity are likely to further reinforce children's information seeking, and to sustain their curiosity so that it can support scientific thinking and learning.

CONCLUSION: A CALL FOR RESEARCH

In this article, I describe evidence from the limited existing research showing that curiosity is important and relates to science learning, and I suggest several mechanisms through which curiosity can support science learning. The general perspective presented here is that science learning can and should be supported by promoting curiosity, and I provide suggestions for promoting (and avoiding the suppression of) curiosity in early childhood. However, much more research is needed to address the complex challenge of educational applications of this work. Specifically, the suggested mechanisms through which curiosity promotes learning need to be studied to tease apart questions of directionality, the influence of related factors such as interest, the impact of context and learning domain on these relations, and the role of individual differences. Both the influence of curiosity on learning and effective ways to promote it likely change in interesting and important ways across development, and research is needed to understand this development-especially through studying change in individuals over time. Finally, it is important to acknowledge that learning does not happen in isolation, and

one's culture and environment have important roles in shaping one's development. Thus, application of research on curiosity and science learning must include studies of the influence of social factors such as socioeconomic status and contexts, the influence of peers, teachers, parents, and others in children's environments, and the many ways that culture may play a role, both in the broad values and beliefs instilled in children and the adults interacting with them, and in the influences of behavior expectations and norms. For example, parents across cultures might respond differently to children's questions, so cross-cultural differences in questions likely indicate something other than differences in curiosity (Ünlütabak et al., 2019). Although curiosity likely promotes science learning across cultures and contexts, the ways in which it does so and effective methods of promoting it may differ, which is an important area for future research to explore. Despite the benefits I present, curiosity seems to be rare or even absent from formal learning contexts (Engel, 2013), even as children show curiosity about things outside of school (Post and Walma van der Molen, 2018). Efforts to promote science learning should focus on the exciting potential for curiosity in supporting children's learning, as promoting young children's curiosity in science can start children on a positive trajectory for later learning.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s) and/or minor(s)' legal guardian/next of kin publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

JJ conceived of the manuscript topic and wrote the manuscript.

FUNDING

This publication was made possible through the support of grants from the John Templeton Foundation, the Spencer Foundation, and the Center for Curriculum Redesign. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the John Templeton Foundation or other funders.

REFERENCES

American Association for the Advancement of Science [AAAS] (1993). *Benchmarks for Science Literacy*. Oxford: Oxford University Press.

Ames, C. (1993). Classrooms: goals, structures, and student motivation. *J. Educ. Psychol.* 84, 261–271. doi: 10.1037/0022-0663.84.3.261

Berlyne, D. E. (1954). An experimental study of human curiosity. *Br. J. Psychol.* 45, 256–265. doi: 10.1111/j.2044-8295.1954.tb01253.x

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., and Schulz, L. (2011). The double-edged sword of pedagogy: instruction limits spontaneous exploration and discovery. *Cognition* 120, 322–330. doi: 10.1016/j.cognition.2010.10.001

Bonwell, C. C., and Eison, J. A. (1991). *Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports. ERIC Clearinghouse on Higher Education*. Washington, DC: The George Washington University.

Chi, M. T. H., Leeuw, N. D., Chiu, M.-H., and Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cogn. Sci.* 18, 439–477. doi: 10.1207/s15516709cog1803_3

Chi, M. T. H., and Wylie, R. (2014). The ICAP framework: linking cognitive engagement to active learning outcomes. *Educ. Psychol.* 49, 219–243. doi: 10.1080/00461520.2014.965823

Chin, C., and Brown, D. E. (2002). Student-generated questions: a meaningful aspect of learning in science. *Int. J. Sci. Educ.* 24, 521–549. doi: 10.1080/09500690110095249

Chin, C., and Osborne, J. (2010). Supporting argumentation through students'. Questions: case studies in science classrooms. *J. Learn. Sci.* 19, 230–284. doi: 10.1080/10508400903530036

Chouinard, M. M., Harris, P. L., and Maratsos, M. P. (2007). Children's questions: a mechanism for cognitive development. *Monogr. Soc. Res. Child Dev.* 72, i–129.

Conati, C., and Carenini, G. (2001). "Generating tailored examples to support learning via self-explanation," in *Proceedings of IJCAI'01, 17th International Joint Conference on Artificial Intelligence*, Seattle, WA, 1301–1306.

Danovitch, J. H., Fisher, M., Schroder, H., Hambrick, D. Z., and Moser, J. (2019). Intelligence and neurophysiological markers of error monitoring relate to Children's intellectual humility. *Child Dev.* 90, 924–939. doi: 10.1111/cdev.12960

Danovitch, J. H., and Mills, C. M. (2018). "Understanding when and how explanation promotes exploration," in *Active Learning from Infancy to Childhood: Social Motivation, Cognition, and Linguistic Mechanisms*, eds M. M. Saylor and P. A. Ganea (Berlin: Springer), 95–112. doi: 10.1007/978-3-319-77182-3_6

Dewey, J. (1910). *How We Think*. Lexington, MA: D.C. Heath and Company. doi: 10.1037/10903-000

Duschl, R. A., and Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Stud. Sci. Educ.* 38, 39–72. doi: 10.1080/03057260208560187

Duschl, R. A., Schweingruber, H. A., Shouse, A. W. (eds) (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8*. Washington, DC: The National Academies Press. doi: 10.17226/11625

Engel, S. (2011). Children's need to know: curiosity in schools. *Harv. Educ. Rev.* 81, 625–645. doi: 10.17763/haer.81.4.h05413131647 3115

Engel, S. (2013). The Case for CURIOSITY. *Educ. Leadersh.* 70, 36–40.

Engel, S., and Labella, M. (2011). Encouraging exploration: the effects of teaching behavior on student expressions of curiosity, as cited in Engel, S. (2011). Children's Need to Know: curiosity in Schools. *Harv. Educ. Rev.* 81, 625–645. doi: 10.17763/haer.81.4.h05413131647 3115

Frazier, B. N., Gelman, S. A., and Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult–child conversation. *Child Dev.* 80, 1592–1611. doi: 10.1111/j.1467-8624.2009.01356.x

French, L. A., Woodring, S. D., and Woodring, S. D. (2013). *Science Education in the Early Years. Handbook of Research on the Education of Young Children*. Available online at: http://www.taylorfrancis.com/ (accessed February 29, 2020).

Gopnik, A. (2012). Scientific thinking in young children: theoretical advances, empirical research, and policy implications. *Science* 337, 1623–1627. doi: 10.1126/science.1223416

Gopnik, A., Griffiths, T. L., and Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Curr. Dir. Psychol. Sci.* 24, 87–92. doi: 10.1177/0963721414556653

Gopnik, A., Meltzoff, A. N., and Kuhl, P. K. (1999). *The Scientist in the Crib: Minds, Brains, and How Children Learn*. New York, NY: William Morrow & Co.

Gordon, G., Breazeal, C., and Engel, S. (2015). Can children catch curiosity from a social robot? *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, 91–98. doi: 10.1145/2696454.2696469

Gottfried, A. E., Preston, K. S. J., Gottfried, A. W., Oliver, P. H., Delany, D. E., and Ibrahim, S. M. (2016). Pathways from parental stimulation of children's curiosity to high school science course accomplishments and science career interest and skill. *Int. J. Sci. Educ.* 38, 1972–1995. doi: 10.1080/09500693.2016.1220690

Harradine, C. C., and Clifford, R. M. (1996). When are children ready for kindergarten? Views of families, kindergarten teachers, and child care providers. *Paper Presented at the Annual Meeting of the American Educational Research Association*, New York, NY.

Howard-Jones, P. A., and Demetriou, S. (2009). Uncertainty and engagement with learning games. *Inst. Sci.* 37, 519–536. doi: 10.1007/s11251-008-9073-6

Heaviside, S., Farris, E., and Carpenter, J. M. (1993). *Public School Kindergarten Teachers' Views on Children's Readiness for School*. US Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Hulme, E., Green, D. T., and Ladd, K. S. (2013). Fostering student engagement by cultivating curiosity: fostering student engagement by cultivating curiosity. *New Dir. Stud. Serv.* 2013, 53–64. doi: 10.1002/ss.20060

Inhelder, B., and Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence: An Essay on the Construction of Formal Operational Structures*. London: Routledge.

Jipson, J. L., Labotka, D., Callanan, M. A., and Gelman, S. A. (2018). "How conversations with parents may help children learn to separate the sheep from the goats (and the Robots)," in *Active Learning from Infancy to Childhood: Social Motivation, Cognition, and Linguistic Mechanisms*, eds M. M. Saylor and P. A. Ganea (Berlin: Springer), 189–212. doi: 10.1007/978-3-319-77182-3_11

Jirout, J., and Klahr, D. (2012). Children's scientific curiosity: in search of an operational definition of an elusive concept. *Dev. Rev.* 32, 125–160. doi: 10.1016/j.dr.2012.04.002

Jirout, J., and Vitiello, V. (2019). "uriosity in the classroom through supportive instruction. *Paper Presented at the SRCD Biennial Meeting*, Baltimore, MD.

Jirout, J., Vitiello, V., and Zumbrunn, S. (2018). "Curiosity in schools," in *The New Science of Curiosity*, ed. G. Gordon (Hauppauge, NY: Nova).

Jirout, J., and Zimmerman, C. (2015). "Development of science process skills in the early childhood years," in *Research in Early Childhood Science Education*, eds K. Cabe Trundle and M. Saçkes (Berlin: Springer), 143–165. doi: 10.1007/978-94-017-9505-0_7

Kelemen, D., Callanan, M. A., Casler, K., and Pérez-Granados, D. R. (2005). Why things happen: teleological explanation in parent-child conversations. *Dev. Psychol.* 41, 251–264. doi: 10.1037/0012-1649.41.1.251

Kidd, C., and Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron* 88, 449–460. doi: 10.1016/j.neuron.2015.09.010

King, A. (1994). Guiding knowledge construction in the classroom: effects of teaching children how to question and how to explain. *Am. Educ. Res. J.* 31, 338–368. doi: 10.2307/1163313

Klahr, D., Matlen, B., and Jirout, J. (2013). "Children as scientific thinkers," in *Handbook of the Psychology of Science*, eds G. Feist and M. Gorman (New York, NY: Springer), 223–248.

Kuhn, D. (2002). "What is scientific thinking, and how does it develop?" in *Blackwell Handbook of Childhood Cognitive Development*, ed. U. Goswami (Oxford: Blackwell Publishing.), 371–393. doi: 10.1002/9780470996652.ch17

Kuhn, D., and Pearsall, S. (2000). Developmental Origins of Scientific Thinking. *J. Cogn. Dev.* 1, 113–129. doi: 10.1207/S15327647JCD0101N_11

Lamnina, M., and Chase, C. C. (2019). Developing a thirst for knowledge: how uncertainty in the classroom influences curiosity, affect, learning, and transfer. *Contemp. Educ. Psychol.* 59:101785. doi: 10.1016/j.cedpsych.2019.101785

Legare, C. H. (2014). The contributions of explanation and exploration to children's scientific reasoning. *Child Dev. Perspect.* 8, 101–106. doi: 10.1111/cdep.12070

Litman, J., Hutchins, T., and Russon, R. (2005). Epistemic curiosity, feeling-of-knowing, and exploratory behaviour. *Cogn. Emot.* 19, 559–582. doi: 10.1080/02699930441000427

Litman, J. A., and Silvia, P. J. (2006). The latent structure of trait curiosity: evidence for interest and deprivation curiosity dimensions. *J. Pers. Assess.* 86, 318–328. doi: 10.1207/s15327752jpa8603_07

Loewenstein, G. (1994). The psychology of curiosity: a review and reinterpretation. *Psychol. Bull.* 116, 75–98. doi: 10.1037/0033-2909.116.1.75

Lowry, N., and Johnson, D. W. (1981). Effects of controversy on epistemic curiosity, achievement, and attitudes. *J. Soc. Psychol.* 115, 31–43. doi: 10.1080/00224545.1981.9711985

Loyens, S. M., and Gijbels, D. (2008). Understanding the effects of constructivist learning environments: introducing a multi-directional approach. *Inst. Sci.* 36, 351–357. doi: 10.1007/s11251-008-9059-4

Luce, M. R., and Hsi, S. (2015). Science-relevant curiosity expression and interest in science: an exploratory study: CURIOSITY AND SCIENCE INTEREST. *Sci. Educ.* 99, 70–97. doi: 10.1002/sce.21144

Martin, A. J. (2011). Courage in the classroom: exploring a new framework predicting academic performance and engagement. *Sch. Psychol. Q.* 26, 145–160. doi: 10.1037/a0023020

Martin, A. J., and Marsh, H. W. (2003). Fear of Failure: Friend or Foe? *Aust. Psychol.* 38, 31–38. doi: 10.1080/00050060310001706997

McDonald, J. P. (1992). *Teaching: Making Sense of an Uncertain Craft*. New York, NY: Teachers College Press.

Metz, K. E. (2004). Children's understanding of scientific inquiry: their conceptualization of uncertainty in investigations of their own design. *Cogn. Instr.* 22, 219–290. doi: 10.1207/s1532690xci2202_3

Mills, C. M., Legare, C. H., Bills, M., and Mejias, C. (2010). Preschoolers use questions as a tool to acquire knowledge from different sources. *J. Cogn. Dev.* 11, 533–560. doi: 10.1080/15248372.2010.516419

Morris, B. J., Croker, S., Masnick, A., and Zimmerman, C. (2012). "The emergence of scientific reasoning," in *Current Topics in Children's Learning and Cognition*, eds H. Kloos, B. J. Morris, and J. L. Amaral (Rijeka: IntechOpen). doi: 10.5772/53885

National Governors Association (2010). *Common Core State Standards*. Washington, DC: National Governors Association.

National Resource Council (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academy Press.

Nelson, D. G. K., Chan, L. E., and Holt, M. B. (2004). When Children Ask, "What Is It? " What Do They Want to Know About Artifacts? *Psychol. Sci.* 15, 384–389. doi: 10.1111/j.0956-7976.2004.00689.x

NPR-Edison Research Spring (2019). *The Smart Audio Report*. Available online at: https://www.nationalpublicmedia.com/uploads/2019/10/The_Smart_Audio_Report_Spring_2019.pdf (accessed February 23, 2020).

Oudeyer, P.-Y., Gottlieb, J., and Lopes, M. (2016). Intrinsic motivation, curiosity, and learning: theory and applications in educational technologies. *Prog. Brain Res.* 229, 257–284. doi: 10.1016/bs.pbr.2016.05.005

Piaget, J. (1926). *The Thought and Language of the Child*. New York, NY: Harcourt, Brace, and Company.

Pianta, R. C., La Paro, K. M., and Hamre, B. K. (2008). *Classroom Assessment Scoring SystemTM: Manual K-3*. Baltimore, MD: Paul H Brookes Publishing.

Post, T., and Walma van der Molen, J. H. (2018). Do children express curiosity at school? Exploring children's experiences of curiosity inside and outside the school context. *Learn. Cult. Soc. Interact.* 18, 60–71. doi: 10.1016/j.lcsi.2018.03.005

Ronfard, S., Zambrana, I. M., Hermansen, T. K., and Kelemen, D. (2018). Question-asking in childhood: a review of the literature and a framework for understanding its development. *Dev. Rev.* 49, 101–120. doi: 10.1016/j.dr.2018.05.002

Rosenshine, B., Meister, C., and Chapman, S. (1996). Teaching students to generate questions: a review of the intervention studies. *Rev. Educ. Res.* 66, 181–221. doi: 10.2307/1170607

Ruggeri, A., and Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition* 143, 203–216. doi: 10.1016/j.cognition.2015.07.004

Saylor, M. M., and Ganea, P. A. (eds) (2018). *Active Learning from Infancy to Childhood: Social Motivation, Cognition, and Linguistic Mechanisms*. Berlin: Springer. doi: 10.1007/978-3-319-77182-3

Schulz, L. E., and Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Dev. Psychol.* 43, 1045–1050. doi: 10.1037/0012-1649.43.4.1045

Schwartz, D. L., and Martin, T. (2004). Inventing to prepare for future learning: the hidden efficiency of encouraging original student production in statistics instruction. *Cogn. Inst.* 22, 129–184. doi: 10.1207/s1532690xci2202_1

Shah, P. E., Weeks, H. M., Richards, B., and Kaciroti, N. (2018). Early childhood curiosity and kindergarten reading and math academic achievement. *Pediatr. Res.* 84, 380–386. doi: 10.1038/s41390-018-0039-3

Turgeon, W. C. (2015). The art and danger of the question: its place within philosophy for children and its philosophical history. *Mind Cult. Act.* 22, 284–298. doi: 10.1080/10749039.2015.1079919

Ünlütabak, B., Nicolopoulou, A., and Aksu-Koç, A. (2019). Questions asked by Turkish preschoolers from middle-SES and low-SES families. *Cogn. Dev.* 52:100802. doi: 10.1016/j.cogdev.2019.100802

van Schijndel, T. J. P., Jansen, B. R. J., and Raijmakers, M. E. J. (2018). Do individual differences in children's curiosity relate to their inquiry-based learning? *Int. J. Sci. Educ.* 40, 996–1015. doi: 10.1080/09500693.2018.1460772

VanLehn, K., Jones, R. M., and Chi, M. T. H. (1992). A model of the self-explanation effect. *J. Learn. Sci.* 2, 1–59. doi: 10.1207/s15327809jls0201_1

West, J., Hausken, E. G., and Collins, M. (1993). *Readiness for Kindergarten: Parent and Teacher Beliefs. Statistics in Brief*. Available online at: https://eric.ed.gov/?id=ED363429 (accessed February 29, 2020).

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Jirout. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 18 August 2020
doi: 10.3389/fpsyg.2020.01718

Family Conversations About Heat and Temperature: Implications for Children's Learning

*Megan R. Luce and Maureen A. Callanan**

Department of Psychology, University of California, Santa Cruz, Santa Cruz, CA, United States

Some science educators claim that children enter science classrooms with a conception of heat considered by physicists to be incorrect and speculate that "misconceptions" may result from the way heat is talked about in everyday language (e.g., Lautrey and Mazens, 2004; Slotta and Chi, 2006). We investigated talk about heat in naturalistic conversation to explore the claim that children often hear heat discussed as a substance rather than as a process, potentially hindering later learning of heat as energy involved in emergent processes. We explored naturalistic speech among children and adults to understand the nature and the frequency of heat- and temperature-related conversations that young children are involved in. This study aims to investigate the actual linguistic resources that children have available as part of a sociocultural approach to cognitive development. Parents' everyday conversations about heat and temperature with their 2–6-year-old children were drawn from the Child Language Data Exchange System (CHILDES) language database and from a parent–child book-reading study. Parents used the word *heat* rarely, but they did so in ways that implied it is a substance. Parents never talked about heat as an emergent process but sometimes as a direct causal process. Most of the heat- and temperature-related talk, however, focused on words like *hot* and *cold* to describe temperature as a property of objects. This investigation of what young children actually experience in everyday conversations is a step toward studying how everyday language may play a role in children's understanding of heat and temperature.

Keywords: science talk, parent–child communication, conceptual change, scientific thinking, sociocultural perspectives

OPEN ACCESS

Edited by:
Ageliki Nicolopoulou, Lehigh University, United States

Reviewed by:
Kathryn Leech, University of North Carolina at Chapel Hill, United States
Andrew Shtulman, Occidental College, United States

***Correspondence:**
Maureen A. Callanan
callanan@ucsc.edu

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *28 February 2020*
Accepted: *23 June 2020*
Published: *18 August 2020*

Citation:
Luce MR and Callanan MA (2020) Family Conversations About Heat and Temperature: Implications for Children's Learning. Front. Psychol. 11:1718. doi: 10.3389/fpsyg.2020.01718

INTRODUCTION

Children develop "intuitive" ideas about key physical concepts and phenomena through everyday navigation of their environments and activities (Piaget, 1974; diSessa, 1996; Wellman and Gelman, 1998; Gopnik and Meltzoff, 2002; Wilkening and Huber, 2004), and yet paradoxically, they often have great difficulty understanding similar concepts later in science classrooms (e.g., McCloskey, 1983). Debates about this discrepancy between "naïve" physics and formal physics have often pitted theories that emphasize the cognitive aspects of conceptual change against theories that emphasize the processes of reasoning occurring in the context of phenomenological experience or sociocultural activities. A more productive path may be to acknowledge that individuals have multiple repertoires (both conceptual and experiential) for understanding complex scientific concepts (Hammer, 2000; Amin, 2009, 2015; Vosniadou, 2009). In this study, we draw upon a sociocultural framework by focusing on a crucial piece that is often left out of these discussions:

a characterization of the everyday experiences within which children come to understand and use such concepts (Rogoff et al., 2018).

Heat and temperature provide an excellent example of a conceptual domain where children's well-developed everyday notions come up against very different scientific notions. For example, toddlers begin to learn about temperature concepts by experiencing and talking about such things as cold water, hot stoves, and warm fuzzy slippers. Despite extensive evidence of young children's clear expression of temperature concepts in everyday language, understanding the concepts of heat and temperature in formal science instruction proves to be difficult for many students even in high school (e.g., Clough and Driver, 1985; Erickson and Tiberghien, 1985; Lewis and Linn, 1994). Some science education researchers have suggested that this results from "misconceptions" in young children's everyday knowledge that are often quite distinct from current "scientific" conceptions of these phenomena (Shtulman, 2017; see Abimbola, 1988; Hammer and Van Zee, 2006, for discussions of the term "misconception"). For example, there is evidence that students often conflate the concepts of heat and temperature (Wiser and Carey, 1983; Wiser, 1995). Chi and colleagues (Slotta et al., 1995; Chi, 2005) argue that students incorrectly conceptualize heat as a substance, which leads to a difficulty in understanding heat as an emergent process.

Prior research and theory suggest the importance of two essential sources of children's learning of concepts such as heat and temperature. First, scientific concepts are argued to emerge in children's phenomenological experiences (diSessa, 1993b). Second, instruction (Chi, 2005), also called "testimony" (Harris, 2002; Harris and Koenig, 2006) or "input" (Gelman, 2009), is argued to serve as another distinct source for children's conceptions. Taking a sociocultural perspective, our view is that these two sources are, in fact, intertwined in the learning that happens in children's everyday lives, where testimony occurs in the midst of phenomenological experience (Rogoff et al., 2018). Several researchers have speculated that everyday language about heat and temperature may mislead children to ideas about these topics that differ from the current scientific ideas. For example, Slotta and Chi (2006) speculate that the ways heat is talked about in everyday contexts may contribute to the miscategorization of heat as a substance rather than as an emergent process. Furthermore, Lautrey and Mazens (2004) suggest that there are few linguistic supports in everyday language for thinking about heat as energy transfer (in contrast to sound, also argued to be an emergent process, for which terms such as *resonate* and *vibrate* are available to intuitively describe it). Despite these speculations, there is very little systematic investigation of the actual everyday linguistic supports for thinking about heat and temperature that are available to young children, either in family settings or in classroom settings. Rather, most of the previous research assesses children's understanding in structured experimental settings (e.g., Wiser and Amin, 2001; Clark and Jorde, 2004). Our goal in this study is to illuminate the potential contribution of spontaneous everyday language during family activity to children's developing conceptions of heat and temperature. Along with others (Rosebery et al., 2010), we argue that children's experiences, interpretations, and meanings of the phenomena of temperature are varied. Therefore, we cannot assume that all children have similar experiences and meanings to draw upon in any given context. In this study, we aim to explore the variation in linguistic contexts in which children converse with others about temperature in order to expand our knowledge of variation and its potential importance in science learning.

As background to the current study, we first present Slotta and Chi's ontological framework for understanding children's concepts of heat. Second, we briefly discuss the alternative views raised by others regarding both children's concepts and scientists' concepts of heat, considering both diSessa's "knowledge-in-pieces" view and the "theory theory" view. Finally, we argue that investigation of parent–child conversations in meaningful activity is necessary to fully understand how children's scientific knowledge develops (Rogoff, 1990; Callanan and Jipson, 2001; Crowley and Jacobs, 2002; Rogoff et al., 2018) and, in this case, to evaluate the claim that everyday language encourages particular misunderstandings. We argue that instead of pitting theories against each other, we can find productive points of overlap that open up new avenues for research.

Slotta and Chi's Ontological View

Slotta, Chi, and colleagues (Slotta et al., 1995; Reiner et al., 2000; Chi, 2005; Slotta and Chi, 2006) proposed that students' difficulty in understanding the nature of heat stems from their misconceptualization of heat as a substance rather than as a process. Drawing on developmental psychology literature (e.g., Keil, 1979, 1983), they have claimed that students incorrectly assign heat to the ontological category of substances. They argue that heat is best conceptualized not as a substance but as an "emergent process," defined as dynamic ongoing interactions of the components within a system, which involves energy transfer and simultaneity at a molecular level (e.g., diffusion of two liquids). More recently, Chi and colleagues have argued that when students do think of heat as a process, rather than as a substance, they still usually misconstrue it as a "direct causal process." In contrast to emergent processes, direct causal processes are defined as simple causal interactions in which the behaviors of the components within a system are distinct and sequential (e.g., a heart pumps, causing blood to circulate). Some argue that the emergent process view of heat is used or preferred by physicists (Slotta et al., 1995; Slotta and Chi, 2006; but see Zemansky, 1970; diSessa, 1993a; Hobson, 1995, for other process views). In one study, physics novices (9th graders) and physics experts (graduate students and postdocs) responded to heat-related problems (e.g., "Two cups of hot coffee are poured into two cups: a Styrofoam cup and a ceramic mug. Both are sealed with airtight lids. After 20 min, what would you predict for the temperatures in both cups?") The participants' language was coded as substance-based if they expressed, for example, that heat gets trapped (contained) or that it moves from object to object, and as emergent process-based if they talked about heat as a process of energy transfer (including excitation of molecules). Physics novices were more likely to use substance predicates than process predicates for the heat problem, whereas physics experts were more likely to use process predicates (Slotta et al., 1995).

Children and non-expert adults in Slotta and Chi's studies responded in ways that suggest they think of heat as a substance

or direct causal process rather than emergent process (Slotta et al., 1995; Reiner et al., 2000; Slotta and Chi, 2006). In support of this view, earlier interview research found that young children (under 6 years) seemed to view heat and cold as distinct substances that are inherent properties of objects and that can be accumulated and contained (Albert, 1978; Erickson, 1979; Clough and Driver, 1985). The oldest children studied (9–10-years-old) mentioned concepts of heat energy transfer similar to conduction (energy transfer due to contact) and to radiation or convection (energy transfer by proximity or currents); however, it is unclear whether these children understood heat as an emergent process.

Chi and colleagues argue that because students do not readily understand the class of emergent processes, they must be instructed to learn this ontological category and reassign the concept of heat to it (Slotta and Chi, 2006). In various studies, they measured which ontology (substance or emergent process) high school and college students used to assign the concept of heat (and other phenomena such as light and electric current) before and after instruction in the emergent process ontology (Slotta et al., 1995; Slotta and Chi, 2006). In one study, instructing undergraduate students in the emergent process ontology facilitated their greater use of emergent process predicates on problems of electric current when compared to a control group (Slotta and Chi, 2006). Wiser and Amin (2001) provided some preliminary evidence for the potential effectiveness of such instruction. Four eighth-grade students were directly instructed that there are two alternative ways to think about heat and temperature: an "everyday" and a "scientific" way. Students were explicitly taught that these ways of thinking make use of the same words (e.g., heat) but convey different meanings (scientific *heat* refers to a process of energy transfer, whereas everyday *heat* can refer to a substance that moves). Children who participated in this metaconceptual teaching unit were argued to have come to view these as compatible but differentiated models based on their tendency to talk about heat as molecular energy when solving problems in a testing situation after the lessons, where the interviewer asked them to clarify their statements in terms of "everyday" and "scientific" heat.

Alternative Views: Knowledge in Pieces and Theory Change

The distinction Slotta and Chi made between substance-based and emergent process-based concepts of heat has been questioned from the standpoint of how both children and practicing physicists think about heat. diSessa (1993a) suggested that people do have everyday experiences with more physics-like ontologies, such as equilibrium changes, and thus may have some intuitive knowledge about processes. He further questions the assumption that most physicists attribute heat, as a form of energy, to an emergent process ontology, arguing instead that physicists may think about energy as a substance-like entity (having location, being conserved, "flowing," even having mass) that is transferred in situations such as heating. In fact, many physicists may, at times, find it appropriate to speak about heat as a substance, as indicated in some physics textbooks (e.g., ter Haar and Wergeland, 1960; Reif, 1965). For example, Reif (1965) mentions that the "amount of heat is transferred" (p. 106) and "heat is absorbed by the system at the lower absolute temperature T and given off by the system." (p. 106). Granting that this language is likely being used in a metaphorical way or as a strategy to avoid the difficulty of explaining the complex non-intuitive notion of emergent processes, the variability in children's, as well as physicists', thinking (or at least ways of expressing concepts) still suggests that knowledge about heat may not be organized strictly in terms of ontological categories such as substance and process (diSessa, 1993b; Clark, 2006; Siegler, 2007). Indeed Chi later revised her claim to include conceptualizations of heat as "direct causal process" as more advanced than a substance-based view, although still inferior to an emergent process view.

Theory change proponents, rather than focusing on children's misconceptions, argue that conceptual change occurs through the gradual formation and the subsequent revision of broad intuitive theoretical frameworks (see Carey, 1985, 1991; Vosniadou and Brewer, 1992; Gopnik and Meltzoff, 2002; Lautrey and Mazens, 2004). While children's statements often seem inaccurate in light of the adult or scientific view, they are argued to be consistent with a coherent alternative theory held by the child. For example, children may first have a theory that conflates heat and temperature, and only later in development separate those into two distinct concepts. Also, children may revise components of theories at different rates, resulting in the differential understanding of two phenomena – sound and heat, for example (Lautrey and Mazens, 2004) – supporting the notion that children can hold distinct theories that can be revised at different rates based on evidence.

Proponents of the knowledge-in-pieces view argue that knowledge about physics is more loosely organized than either ontological categories or theories, such that children, as well as physicists, do not have a single conception of heat or may necessarily use several ontologies when thinking about heat (diSessa, 1988, 1993a; Gupta et al., 2010). On this view, the expression of knowledge may appear as assemblies of pieces coming together in the moment to solve a particular problem or as a process of drawing upon various resources one has for thinking about a phenomenon (Hammer, 2000). Findings that are argued to support this view include those showing that people are likely to reason differently across contexts (Clark and Jorde, 2004; Clark, 2006) and that thinking about varied prior experiences can help students engage with different aspects of a phenomenon (Rosebery et al., 2010). For example, when responding to heat-related word problems, eighth-grade students often expressed multiple contradictory ideas, which included claiming at one point that metal objects will be of the same temperature as glass objects in an oven and claiming at another point that the metal objects will be hotter because metal is a good conductor (Clark, 2006). Children also reasoned differently depending on the context of the task and used prior experiences to support their reasoning. For example, one student claimed that wood and metal objects would be of the same temperature when placed in a hot trunk but not when placed in a hot oven. Others claimed that metal is a good insulator because people wrap soda cans in aluminum foil to keep them cold (see also Lewis and Linn, 1994; Brookes and Etkina, 2015).

The literature presents conflicting results regarding how children's knowledge is organized and how learning occurs, and multiple theories have different explanations for the same findings. The available evidence has been argued to support the ontological category view, the theory theory view, and the knowledge-in-pieces view (diSessa, 1988; Lautrey and Mazens, 2004; Slotta and Chi, 2006), and yet theorists on all sides of the debate have competing explanations for similar data. For example, while some theorists see inconsistent responses about heat as evidence of "knowledge in pieces," others argue that they are evidence that children are in transition between two coherent theories (Lautrey and Mazens, 2004) or that children have a coherent but incomplete version of an ontological view (Chi, 2005).

Studies of children's conceptions of phenomena such as heat and sound (e.g., Mazens and Lautrey, 2003; Clark, 2006) have aimed to test a theory theory view against a "knowledge-in-pieces" view. By using semi-structured interviews to assess 6–10-years-old children's predictions and explanations for sound-related problems, Mazens and Lautrey (2003) found that children were not consistent in interpreting sound as a substance. Children instead assigned some aspects of sound to process ontology, suggesting that they entertain multiple ontologies for sound at the same time. Mazens and Lautrey (2003) interpreted this as evidence for theory theory and as evidence for the gradual transition from one intuitive theory to another. However, one might argue instead that such variation in the same child's responses could support a more flexible view of knowledge, such as the knowledge-in-pieces view (Gupta et al., 2010). Furthermore, Slotta and Chi (2006) revised their claim that the substance ontology must be replaced by an emergent process ontology, suggesting that these two ontological categories can exist in parallel, where students learn which one is appropriate in which situations. Thus, it is difficult to empirically distinguish among these three theoretical approaches. Indeed more recent work argues for developing a new view that moves beyond trying to choose among theories and instead considers how learners incorporate multiple types of resources (including embodied experience, conceptual metaphors and other cognitive models, scientific language and everyday language) in understanding abstract scientific concepts (Hammer, 2000; Amin, 2015).

In order to integrate the sources available to the child, one type of data that is largely missing is the detailed analysis of young children's exposure to ideas about heat and temperature as they relate to everyday discourse. Regardless of one's theoretical orientation, we suggest that a crucial next step is to investigate the varied everyday situations in which young children experience talk and action related to heat and temperature. In particular, we need more information about everyday discourse in order to articulate the data available to children as they create and revise their understanding of heat and temperature across contexts.

Social Context of Scientific Thinking

From each of the outlined perspectives – ontological, knowledge-in-pieces, theory change – the way heat is talked about in everyday contexts is a potentially important piece of the puzzle regarding how children come to understand heat and respond to heat-related problems across contexts. Slotta and Chi and others speculated that commitments to ontologies such as substance may be reinforced by everyday language (Zemansky, 1970; Erickson and Tiberghien, 1985; Hobson, 1995; Slotta and Chi, 2006), and recent data support some of these claims (Brookes and Etkina, 2015). For example, someone might say: "Shut the door, you're letting out all of the heat." Hearing *heat* used as a noun with the implication that it is a substance that can move from one location to another may present an obstacle to understanding heat as energy or as an emergent process.

While diSessa (2000) and others would disagree that speaking about heat as if it is a substance or a direct causal process is categorically limiting or wrong, the knowledge-in-pieces view also calls for an account of how children encounter discourse about heat and temperature in everyday situations. diSessa's view focuses mostly on the phenomenological experiences of heat and temperature, but we suspect that he would agree that it is important to investigate how language is used to describe or make sense of those experiences as well. diSessa (2000) describes the potential disconnect between learning the words and the meanings of a physical phenomenon and how that phenomenon is experienced. Related to this point, a recent study by Jeppson et al. (2017) used infrared cameras to support 7–8-year-old students in discussing and conceptualizing heat as a process in the context of their embodied experience. Thus, it seems important to investigate how students merge experiences of heat and temperature with words and meanings that people use to communicate about them (see also Rosebery et al., 2010).

Finally, from a theory theory perspective, Lautrey and Mazens (2004) argue that children's theories of heat and sound may be partly influenced by ways that language does or does not provide support for thinking about such difficult concepts. For example, hearing the expressions *vibrate* and *resonate* for the process of sound may provide children with a "direct intuition" about the mechanism of sound transfer. They suggest that language supports for heat transfer seem less available than those for sound in everyday experiences. This points to the importance of considering everyday experiences in children's theory revision processes.

We argue, then, that what is sorely needed in the field is an analysis of the language that young children actually hear from their parents (and other adults) about heat and the related concept of temperature. Do parents talk about heat in ways that suggest that heat is an entity or a substance that moves from one object to another? How do parents and children talk about the related concept of temperature? Parent–child conversation is a prominent social interaction context for young children. Research on Western parent–child conversations reveals that such contexts can offer rich opportunities for conceptual development and scientific thinking in which parents help children understand the world as they experience it and offer knowledge about the world to which children may not have direct access (Vygotsky, 1978; Bruner, 1983; Tizard and Hughes, 1984; Callanan and Oakes, 1992; Snow and Kurland, 1996; Crowley et al., 2001; Harris, 2002). Investigating the linguistic context in which children learn about heat and temperature would further

allow for future analysis of the links between language and non-linguistic conceptualizations of heat. We can also ask about whether parents' talk about heat and temperature varied as a function of children's age within the age range from toddlers to school entry. It is possible that parents' language may change over this age range as they perceive children to become more sophisticated in their comprehension.

The present study investigates children's early experiences with heat- and temperature-related concepts in several corpora of naturalistic parent–child conversations (three cross-sectional and three longitudinal data sets). Conversations were drawn from the Child Language Data Exchange System (CHILDES) and from a study of video-recorded book-reading sessions (Callanan et al., 1995) to explore the range of everyday contexts in which parents and children talk about heat and temperature.

MATERIALS AND METHODS

CHILDES Database

We used the CHILDES as a source of naturalistic family conversation about heat and temperature. Created by Brian MacWhinney (1995), CHILDES is a shared database of transcripts and audio files of children's recorded conversation and speech. The corpus consists of Institutional Review Board-approved data that are contributed by researchers. For this study, British English conversations were excluded because of our unfamiliarity with the dialect. Four criteria were used to select databases to include in this study: The children studied were between the ages of 2 and 7 years, the recordings of speech were in everyday settings such as homes, parents were present during the recordings, and the database included only one study (some databases include transcripts from multiple studies with different samples of children and different settings).

Participants

The CHILDES data were drawn from five databases: Hall (Hall and Tirre, 1979), Gleason (Gleason et al., 1984), Sachs (Sachs, 1983), Kuczaj (Kuczaj, 1976), and Brown (Brown, 1973). **Table 1** displays information about the nature of the data collection, demographic information, age, and gender for the participants.

Materials and Procedure

The speech recordings were transcribed in the CHAT transcription format and uploaded to the CHILDES database by each contributor (MacWhinney, 1995). We used the CLAN program (*kwal* command), designed for searching and analyzing transcript information, to search the transcripts for the heat- and temperature-related words shown in **Table 2**. The keyword search was broadened beyond the term *heat* because it was predicted that parents and children might use many other words conceptually related to heat, such as *hot*, *cold*, and *warm*.

The utterance that contained the keyword and the three lines before and after the utterance were returned in CHAT format. These seven-line conversation segments were then coded. The entire transcript was consulted if more conversational context was necessary to interpret the utterance. Examples of seven-line conversation segments are shown below:

Keyword: hot. File "gas.cha": line 14452
Mother: you observe more than I do #.
Child: what is dis [: this] black stuff?
Mother: that's just the cheese browning.

TABLE 1 | Participant information for each data source.

Data source	Data collection	Demographics	Age and gender
CHILDES			
Hall	Study of vocabulary use in children from various socio-economic and racial groups. Families observed at home, school, and in route to school.	39 children sampled from 4 populations: Black working-class White working-class Black middle-class White middle-class	4;6–5;0 Gender N/A (roughly equal boys and girls)
Gleason	Study of acquisition of communicative competence. Dinnertime conversations between children, mothers, and fathers in their homes.	22 children White, middle-class Boston, MA area	2;0–5;2 11 girls 11 boys
Sachs	Researcher collected speech samples from her daughter in her home. Longitudinal data. (Naomi)	1 child	1;2–5;1 girl
Kuczaj	Researcher collected speech samples from his son in his home. Longitudinal data. (Abe)	1 child	2;4–5;0 boy
Brown			
Adam	Speech collected from child in his home.	1 child Middle-class, Black family	2;3–4;10 boy
Eve	Speech collected from child in her home.	1 child Middle-class White family	1;6–2;3 girl
Sarah	Speech collected from child in her home.	1 child Working-class White family	2;3–5;1 girl
Snowman Study	Videorecorded interactions of parents and children reading a book together in their home.	51 children Mostly middle- to upper-middle-class White families.	2;0–5;3 25 girls 26 boys

TABLE 2 | The heat- and temperature-related words (and frequency of occurrence) searched for in the Child Language Data Exchange System databases and Snowman Study transcripts.

Cold (931)	Cooler (2)	Hotter (7)	Warmest (0)
Colder (5)	Cooling (2)	Hottest (0)	Warming (3)
Coldest (4)	Heat (42)	Warm (257)	Warms (3)
Cool (126)	Heated (4)	Warmed (8)	
Cooled (8)	Hot (818)	Warmer (14)	

Mother: when the oven is hot, cheese browns a little bit.
Mother: it taste good.
Mother: sure you don't want a taste #?
Child: you please don't have any of moi [: my] stuff.

Keyword: warm. File "ded.cha": line 690
Father: it's not time yet.
Father: it's not time to go yet.
Child: yeah well I'm getting in [!].
Father: no, because you will be too warm in this inside.
Child: no I won't.
Father: wait Mommy still has a little more time left, you're gonna [: going to] be too warm and then you'll go out and get a cold.
Father: right?

Keyword: warm. File "adam03.cha": line 1495
Child: Adam don't wear wear shoe.
Mother: yes # Adam does wear shoes.
Child: take shoe off?
Mother: shoes help keep your feet warm.
Child: keep feet warm?
Mother: what do you wear over your shoes when it's raining?
Child: oh no xxx wear shoes.

Coding Heat and Temperature Words

A five-category coding scheme was developed to capture the ways that parents and children talked about the word *heat* and other temperature words: *substance*, *emergent process*, *vague process*, *property*, and *other meaning*. **Table 3** shows a list of the coding categories and examples from the data. The coding scheme began with the original substance and emergent process codes[1] of Slotta et al. (1995), and other codes were added.

The *substance* code was used to identify cases in which heat or heat- and temperature-related ideas were expressed in substance predicates, implying that heat can be contained, quantified, absorbed, and accumulated (e.g., "You can't feel the heat?" "The heat. . .is out in the room," "Heat right in the fire," "Close that window, it's cold in there we got no heat").

The *emergent process* code was used to identify cases in which heat or heat- and temperature-related ideas were expressed with the particular process predicates that Slotta et al. (1995) identify as reflecting a conceptualization of heat as an acausal interaction or a constraint-based interaction. Emergent process predicates imply that heat is a process involving transfer, excitation, interaction, equilibrium-seeking, simultaneity, and uniform and continuous interactions of components (Chi, 2005; see **Table 3** for an example).

TABLE 3 | Heat- and temperature-related coding categories and examples from the data.

Type of talk	Examples
Substance	We don't have that kind of heat. So it must be very sensitive to heat. Child: I see fire. Parent: What about the heat? He likes the cold.
Emergent process	The stove transferred the heat to the pot. (Slotta et al., 1995)
Vague process	I'm getting warmer. Okay, I'll put the roast back in to keep it warm. To heat up the house.
Proximity	(he's too near the fire) he's going to get too hot. You getting hot in here?
Time	Let it warm up first. Oh, it's cool now?
Contact/Presence	That's a jacket to keep me warm. The sun makes you hot, right?
Property	Child: It's cold. Parent: What's cold? Your milk? He's hot but I'm cold. The sun's hot, right.
Other	Did you catch a cold? I'm hot on his trail.

The *vague process* category was added because parents and children sometimes identified temperature change as a process, without including the word heat. They discussed temperature change in ways that did not reflect an emergent process conceptualization (at the molecular level) and instead focused on processes at a more macroscopic level. Vague process includes mentions of the heating or temperature change process without indicating heat as a substance (e.g., "Just heat it and serve it," "We have to wait for it to heat up," "It's getting warmer," "You don't want to get cold"), suggesting that heating and cooling are macro-level processes. References to temperature change were also further coded into three sub-categories: *proximity* indicated that temperature change is related to whether an object is close to or far from a heat source or indicated the location of an object, *time* included explicit mentions of time, and *contact* indicated that temperature change is associated with coming into contact with or being in the presence of a heat source (see **Table 3**).

In addition to the substance and process codes, a *property* code was developed to capture talk about heat and temperature words as adjectives, suggesting that they are properties of objects (e.g., "That's a hot fire," "My peas are cold").

The *other meaning* category was included to capture metaphorical and other uses of hot and cold as well as isolated words that did not have sufficient context to determine the intended meaning (see **Table 3**).

The utterance that contained the keyword was the unit of analysis, and utterances that contained more than one keyword were coded only once. For example, the utterance *this is hot and that is cold* was given only the property code. This was done to keep the coding unit at the utterance level already

[1]The term emergent process was introduced later; the early work referenced constraint-based interactions.

determined in the CHAT transcription guidelines. For utterances that contained two keywords that could be coded differently, the primary coder determined which code best captured the meaning of the conversation within which the utterance was embedded. Inter-coder reliability was established on 20% of the utterances. Percent agreement ranged from 85 to 89% and Cohen's Kappas ranged from 0.75 to 0.80 (excellent level of agreement according to Bakeman and Gottman, 1986). Disagreements were resolved through discussion.

Coding Conversational Contexts

A nine-category coding scheme was developed to capture the range of conversational contexts in which parents and children talked about aspects of heat and temperature. Each of the 1,738 heat- and temperature-related utterances in our selected CHILDES data was coded into one of the nine categories. The categories aim to capture context in a broad sense that considers both activity setting and topic discussed, including either participating in or talking about meal times, weather, body temperatures, dressing in clothing, touching objects, heat sources, bathing, and swimming. A ninth category (other) included the idiomatic uses of the keywords as well as cases where the context was not clear from the transcript. To establish inter-coder reliability, two people coded 33% of the utterances. Cohen's Kappa was 0.95 (excellent agreement). Disagreements were resolved through discussion.

Snowman Book-Reading Task

Participants

Fifty-one parent–child dyads participated in this book-reading study (see Callanan et al., 1995), including four age groups: 2-year-olds (six boys and six girls, mean age = 23.87 months, *SD* = 1.70, range = 22–27 months), 3-year-olds (seven boys and seven girls, mean age = 35.71 months, *SD* = 1.43, range = 33–38 months), 4-year-olds (six boys and six girls, mean age = 45.58 months, *SD* = 2.71, range = 42–50 months), and 5-year-olds (seven boys and six girls, mean age = 58.46 months, *SD* = 1.94, range = 56–63 months). Of the 51 parents, 47 were mothers and four were fathers. The families were recruited through daycare centers and informal contacts of the researchers and other participants (see **Table 1** for data collection and demographic information).

Materials and Procedure

Parents and children read the wordless children's book *The Snowman* (Briggs, 1978). The storybook depicts a boy's adventure in making a snowman that "comes to life" and is introduced to household items such as a fireplace, stove, hot water, and a freezer. Because the book does not contain words, parents and children may narrate the story as they wish. It was expected that because of the snowman's encounters with hot and cold household items and his propensity to melt, parents and children would talk about heat and temperature.

The researchers video-recorded the families in their homes. The parent–child dyads (one parent and one child) first engaged in making muffins together, with the exception of the 2-year-olds who only read the book. While the muffins were baking, the parent and the child read the book together. The video-recorded sessions were transcribed.

Coding

The transcripts were searched for temperature-related words (see **Table 2**) as in the CHILDES procedure. The five-category coding scheme (substance, emergent process, vague process, property, and other) described in the CHILDES coding section above and in **Table 3** was also used for the Snowman Study analysis. Inter-coder reliability was established on 20% of the utterances. Percent agreement was 89–91% and Cohen's Kappa was 0.83 (excellent; Bakeman and Gottman, 1986). Disagreements were resolved through discussion. Because in this data set each parent-child dyad was engaged in the activity of book reading, the conversation context coding scheme was not relevant.

RESULTS AND DISCUSSION

This section is divided into three main analyses (with discussion following each analysis) that include both the CHILDES and the Snowman Study, and main findings are provided and discussed for each analysis. Analysis 1 tests the claim that children hear adults talk about *heat* as a substance. Analysis 2 explores the range of different heat- and temperature-related words parents and children used, asking how often they occurred in everyday conversations, and how often they were used as property, substance, or process. Analysis 3 describes how such words were used across different everyday activity contexts for the CHILDES databases only.

Analysis 1: The Keyword Heat Only

Because the word *heat* is of particular importance in conceptions of heat and temperature, it was separated out from all other keywords for analysis. This analysis tested the claim that young children often hear the word *heat* talked about as a substance (Slotta and Chi, 2006).

Main Finding 1

Parents used the word *heat* rarely, but when they did, they most often talked about heat in ways that are consistent with it being a substance.

Of the 2,138 coded utterances across all CHILDES databases and the Snowman Study data, 49 contained the keyword *heat*. Of these instances, 42 (86%) were spoken by parents and seven (14%) were spoken by children. For children, three (43%) of their references to heat were coded as substance, three (43%) as vague process, and one (14%) as other. For parents, 29 (69%) of their references to heat were coded as substance and 13 (31%) were coded as vague process [$\chi^2(1, N = 42) = 6.10$, $p = 0.014$]. As Chi and colleagues would predict, emergent processes were never discussed.

As an exploratory analysis, we developed a sub-category of the substance code (substance-involved-in-direct-process) that was used to identify the instances when parents or children seemed to identify heat as a substance but also implied that it was involved in a direct causal process where, based on Chi's

conceptualization, there is directionality and the movement is not described explicitly as continuous or simultaneous (e.g., "This is where the heat comes in this house," "And they're still sending up the heat," "You hold it in your hands and it gets soft so that you get the heat from your hands")[2]. The percent agreement between two coders was 88%; Cohen's Kappa = 0.80 for all *heat* statements.

Further analysis including the substance-involved-in-direct-process code made little difference in the pattern of children's talk; of their three substance comments, one was recoded as substance-involved-in-direct-process. For parents, 20 of their 29 references to heat remained coded as substance, and nine were re-coded as substance-involved-in-direct-process. Thus, a strict interpretation of Chi's idea is supported – parents do most often talk about heat in ways that are consistent with it being a substance and never as an emergent process. Children were also heard talking about heat involved in macro-level processes. Almost equal in frequency to the talk about heat as a substance were the comments about heat as a general process, if we combine substance-involved-in-direct-process and vague process talk compared with substance talk (52 versus 48%). However, it is important to emphasize that this is very different from emergent process at the molecular level. When parents used the word *heat*, then it was most often either substance-based or focused on direct (macro-level) causal processes. At the same time, the word *heat* occurred rarely in these conversations, which suggests that children may actually be getting very few opportunities to hear about how adults conceptualize the concept of heat in everyday talk. In the few instances where children used the word *heat*, they were equally likely to use substance-based language and vague process-based language.

Most of the conversations about heat and temperature did not include the word *heat*. The next analysis focuses on utterances that contained the other temperature-related words.

Analysis 2: All Temperature Keywords (Excluding Heat)

Cold and *hot* were the most frequently observed words (see **Table 2**). **Table 4** displays the percentage of utterances in each database containing temperature words that were coded as each type of talk. Heat- and temperature-related words were never talked about as emergent processes in any CHILDES database or in the Snowman Study; therefore, the category was excluded from analyses. Because the metaphorical or other meanings of heat- and temperature-related words were not of interest for the current study, the other category was also excluded from subsequent analyses. In databases where children's gender was known, analyses were initially conducted with gender as a between-subjects factor. No significant effects or interactions involving gender were found; therefore, it was excluded from all subsequent analyses. All analyses were conducted separately for parents and children.

TABLE 4 | Percentage of "heat" and temperature utterances coded as each type of talk by database.

	Abe	Adam	Eve	Gleason	Hall	Naomi	Sarah	Snowman
Utterances (n)	289	127	130	127	592	155	318	400
Property	78%	66%	68%	83%	68%	78%	76%	69%
Substance	< 1%	< 1%	6%	0	3%	1%	4%	5%
Emergent Process	0	0	0	0	0	0	0	0
Vague Process	19%	17%	23%	14%	18%	12%	8%	24%
Other	4%	17%	2%	3%	10%	8%	12%	3%

In analysis 2, our second and third main findings are stated and then the corresponding analyses from each of the relevant databases are presented.

Main Finding 2

Both parents and children talked about heat- and temperature-related words as properties of objects more often than they talked about them as vague processes or as substances in all but one of the databases. The only exception is that parents of older children in one database (Gleason) talked equally about properties and vague processes.

Main Finding 3

Although the parents did not talk in ways that are consistent with emergent processes or substances, they did sometimes give information about vague processes. In addition, parents in the Snowman Study mentioned three aspects of the process of temperature change.

CHILDES Database: Gleason Dinner

In this database, neither parents nor children talked about temperature-related words as substances. Therefore, a paired-samples *t*-test compared the parents' use of two types of talk: property and vague process. As shown in **Figure 1**, the parents talked about temperature-related words as properties of objects (M = 3.36, SD = 5.57) more frequently than they talked about them as vague processes [M = 0.82, SD = 1.59; $t(21) = 2.73$, $p < 0.02$, $d = 0.62$].

The 18 children whose ages were known were divided into two age groups: younger children who were 2–3.5 years old (four girls and five boys, mean age = 32 months, SD = 4.50) and older children who were 3.5–5 years old (three girls and six boys, mean age = 50.67 months, SD = 5.19). An analysis of how parents' talk about heat and temperature varied with age was conducted in a 2 (type of talk: property, vague process) × 2 (age group: younger, older) mixed ANOVA on the frequency of the parents' use of these types of talk. Type of talk was a within-subject factor and age group was a between-subjects factor. In a significant interaction [$F(1, 16) = 7.00$, $p < 0.02$, $\eta_p^2 = 0.30$], the parents of younger children used heat- and temperature-related words as properties of objects (M = 7.00, SD = 7.21) more than as vague processes [M = 1.33, SD = 2.18; $t(8) = 3.18$, $p = 0.013$, $d = 1.06$], whereas

[2]The substance-involved-in-direct-process coding category was applied to the heat- and temperature-related words that were originally coded as substance. This analysis of the few instances of temperature words used as if they were substances revealed that the coding category did not apply conceptually. For example, "He's standing out in the cold" and "Because snowmen love the cold" seemed to reflect a substance-based conceptualization of coldness (the idea of coldness is a misconception in itself) but did not implicate cold(ness) in a process as was done with *heat*.

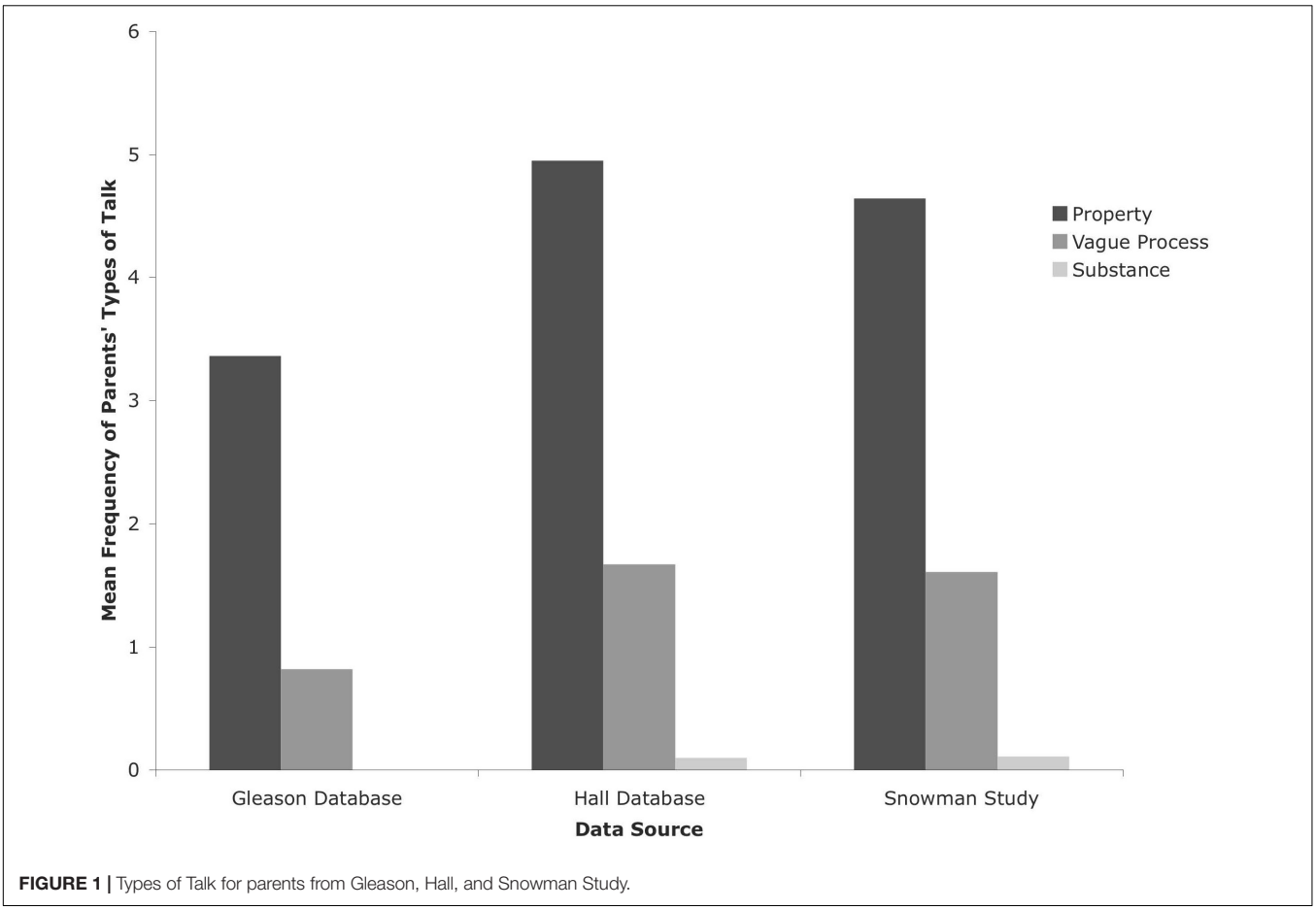


FIGURE 1 | Types of Talk for parents from Gleason, Hall, and Snowman Study.

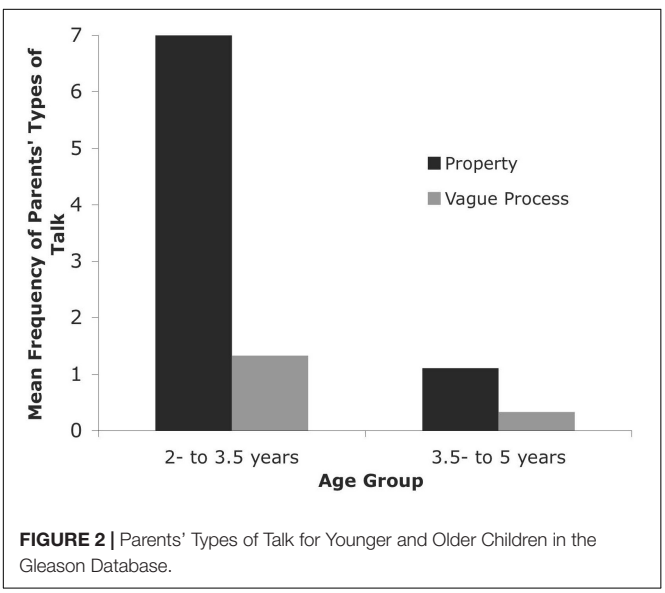


FIGURE 2 | Parents' Types of Talk for Younger and Older Children in the Gleason Database.

the parents of older children talked about properties (M = 1.11, SD = 1.96) and vague processes (M = 0.33, SD = 0.71) with similar frequency [$t(8) = 1.58$, $p = 0.15$] (see **Figure 2**).

The children's use of the two types of talk was compared in a paired-samples t-test. The children talked about heat- and temperature-related words as properties (M = 1.00, SD = 1.66) more frequently than they talked about them as vague processes [M = 0.23, SD = 0.61; $t(21) = 2.85$, $p < 0.01$, $d = 0.62$]. In a separate mixed ANOVA on younger and older children's talk, no age difference was found.

CHILDES Database: Hall

The parents' use of the types of talk (property, substance, and process) was analyzed in a one-way repeated-measures ANOVA, revealing a significant type of talk effect [$F(2, 37) = 17.29$, $p < 0.001$, $\eta_p^2 = 0.48$], as shown in **Figure 1**. Pairwise comparisons indicated that parents talked about heat- and temperature-related words as properties (M = 4.95, SD = 5.37) more frequently than they talked about them as substances [$M = 0.10$, $SD = 0.38$; $t(38) = 5.68$, $p < 0.001$, $d = 1.27$] or as vague processes [$M = 1.67$, $SD = 1.88$; $t(38) = 4.75$, $p < 0.001$, $d = 0.82$] and that parents talked about heat- and temperature-related words as vague processes more frequently than as substances [$t(38) = 5.09$, $p < 0.001$, $d = 1.16$].

The categories of children's talk were compared using one-way repeated-measures ANOVA. Type of talk was again significant [$F(2, 37) = 24.27$, $p < 0.001$, $\eta_p^2 = 0.57$]. Pairwise comparisons

indicated that children talked about heat- and temperature-related words as properties ($M = 5.77$, $SD = 5.25$) more frequently than as substances [$M = 0.10$, $SD = 0.38$; $t(38) = 6.87$, $p < 0.001$, $d = 1.52$] or vague processes [$M = 0.92$, $SD = 1.27$; $t(38) = 6.24$, $p < 0.001$, $d = 1.27$] and that children talked about heat- and temperature-related words as vague processes more frequently than as substances [$t(38) = 3.97$, $p < 0.01$, $d = 0.88$]. Both parents and children had similar patterns of using property-, substance-, and vague process-based language with heat- and temperature-related words.

Snowman Study Data

As with the CHILDES data, we analyzed parents' talk in the Snowman book-reading study with a 3 (type of talk: property, substance, vague process) × 4 (age: 2, 3, 4, 5-years-old) mixed ANOVA with type of talk as a within-subject factor and age as a between-subjects factor. The parents' types of talk varied systematically [$F(2, 46) = 49.89$, $p < 0.001$, $\eta_p^2 = 0.68$], as shown in **Figure 1**. The parents talked about heat- and temperature-related words as properties ($M = 4.64$, SD = 3.55) more frequently than as substances [$M = 0.11$, $SD = 0.38$; $t(50) = 9.12$, $p < 0.001$, $d = 2.15$] or vague processes [$M = 1.61$, $SD = 1.45$; $t(50) = 6.44$, $p < 0.001$, $d = 1.12$] and as vague processes more frequently than as substances [$t(50) = 7.08$, $p < 0.001$, $d = 1.42$].

The children's uses of these types of talk were analyzed in a 3 (type of talk: property, substance, vague process) × 4 (age: 2, 3, 4, 5-years-old) mixed ANOVA. A significant main effect of type of talk [$F(2, 46) = 7.36$, $p < 0.01$, $\eta_p^2 = 0.24$] was followed up by pairwise comparisons showing that children talked about heat- and temperature-related words as properties ($M = 0.53$, $SD = 0.94$) more frequently than as substances [$M = 0.02$, $SD = 0.14$; $t(50) = 3.74$ $p = 0.001$, $d = 0.76$] or as vague processes [$M = 0.06$, $SD = 0.22$; $t(50) = 3.73$, $p < 0.01$, $d = 0.70$]. Furthermore, a main effect of age [$F(3, 48) = 4.44$, $p < 0.01$, $\eta_p^2 = 0.22$] showed that 5-year-olds produced more heat- and temperature-related words ($M = 0.46$, SD = 0.31) than did 2-year-olds [$M = 0.05$, $SD = 0.32$; $t(23) = 3.15$, $p < 0.02$, $d = 1.31$] or 3-year-olds [$M = 0.07$, $SD = 0.32$; $t(25) = 3.15$, $p < 0.02$, $d = 1.21$], likely reflecting the older children's greater productive vocabulary or contribution to conversations. The children's types of talk did not vary by age group.

Vague process sub-codes

The parents in the Snowman Study also made some references to the three aspects of temperature change (proximity, time, and contact) (Parents and children in the CHILDES databases made references to these three aspects too infrequently to run analyses.) For example, a parent made reference to proximity by saying, "Cause you know what happens to snow when it gets *near* something hot?" Another parent made reference to contact by saying, "(eating ice cubes) makes him good and cold." Some parents also made reference to time in the Snowman book, "And he goes out to play again. *Now* he's all warm." The higher frequency of these comments in the Snowman book may result from its focus on situations that are explicitly about temperature change and impending melting.

To examine parents' talk about these three specific aspects of temperature change in the Snowman book study, a 3 (vague process sub-codes: proximity, time, contact) × 4 (age: 2-, 3-, 4-, 5-year-olds) mixed ANOVA was conducted on the frequency of parents' use of these types of talk. Vague process sub-code was a within-subject factor and age was a between-subjects factor. In a main effect of vague process sub-codes [$F(2, 46) = 4.57$, $p < 0.02$, $\eta_p^2 = 0.17$], the parents made significantly more references to contact ($M = 0.37$, $SD = 0.59$) than to proximity [$M = 0.10$, $SD = 0.36$; $t(50) = 2.64$, $p = 0.03$, $d = 0.55$] or to time [$M = 0.10$, SD = 0.36; $t(50) = 2.98$ $p < 0.02$, $d = 0.56$]. There was no main effect of age and no interaction.

CHILDES Databases: Brown, Kuczaj, and Sachs

The longitudinal data from the five children in these databases (Abe, Adam, Eve, Naomi, and Sarah) are presented in **Figures 3–7**. The frequency with which the heat codes were observed in the parents' speech is presented for each child by age grouping. The patterns in each longitudinal sample are case study examples that

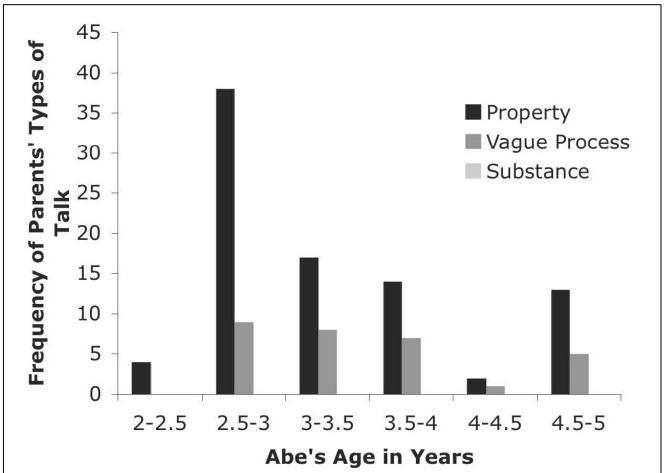


FIGURE 3 | Types of Talk for Parents from CHILDES Abe Database.

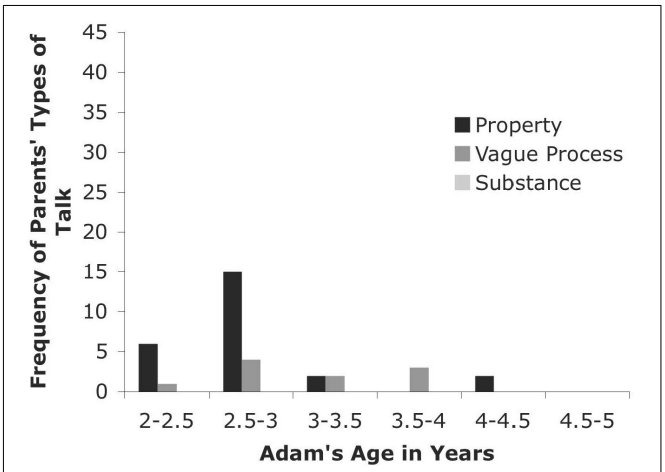


FIGURE 4 | Types of Talk for Parents from CHILDES Adam Database.

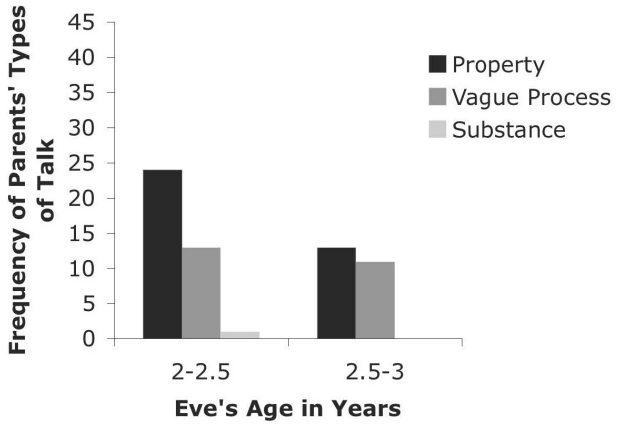


FIGURE 5 | Types of Talk for Parents from CHILDES Eve Database.

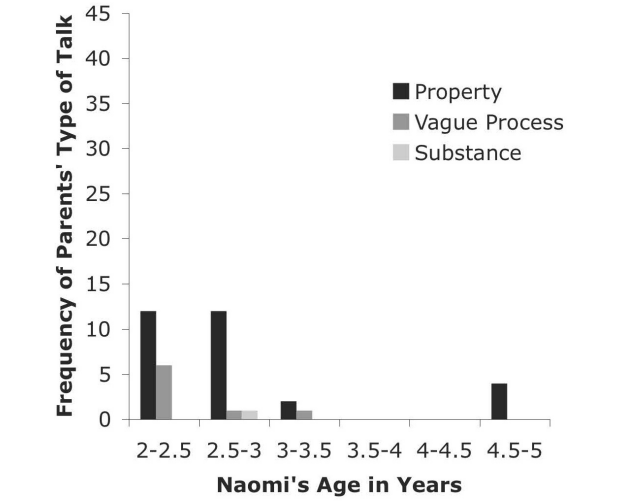


FIGURE 6 | Types of Talk for Parents from CHILDES Naomi Database.

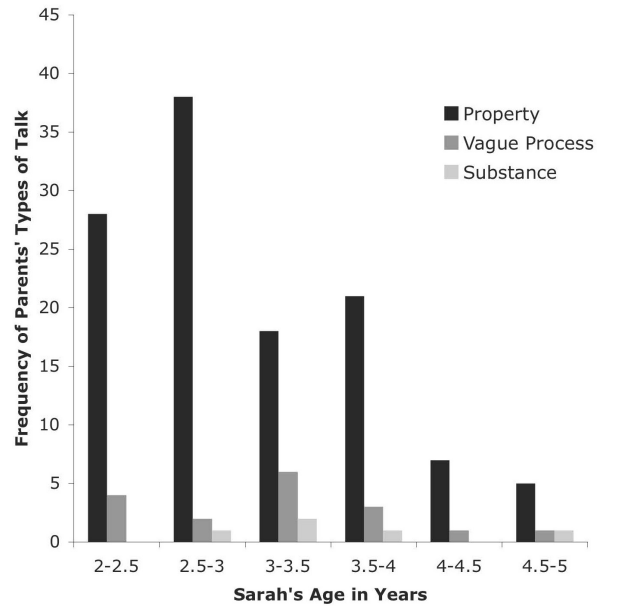


FIGURE 7 | Types of Talk for Parents from CHILDES Sarah Database.

support the cross-sectional results. In general, the parents talked about these words as properties more often than as a substance or as a vague process. Children's talk in these samples (although not depicted) followed a similar pattern.

In summary, across these three cross-sectional data sets and three longitudinal data sets, parents talked about heat- and temperature-related words as properties of objects more often than they talked about them as vague processes or as substances. A question to ask is whether this property-based talk might be related to a substance or a process view of heat. On the one hand, one could argue that thinking of heat as a substance might support the idea that heat is a property of hot objects. Parents in the Gleason Dinner database talked about temperature as property more with 2–3.5-year-olds but talked about properties and processes similarly with 3.5–5-year-olds. This finding is interesting in light of the evidence that young preschoolers are likely to answer heat- and temperature-related questions as if they believe that "hot" and "cold" are intrinsic properties of objects (e.g., Albert, 1978). On the other hand, one might argue that property talk could support a macro-level process view (although not a molecular emergent process view), depending on the real-world experiences within which children are hearing this language. In particular, if parents label objects that are not stable in their "hotness" or "coldness," children might experience changes in objects' temperature and see that the words "hot" and "cold" refer to transitory states. While this would still not be an emergent process view of heat at the molecular level, it could be different from the concept of heat as a substance.

As an exploratory analysis, we further coded all *property* statements made by parents as to whether they referred to items that generally have stable temperatures (including ice, fire, the sun, winter, freezers, etc.) or to items for which temperature is transitory (all other references). The percent agreement between two coders was 87% for 20% of parents' property statements (Cohen's Kappa = 0.66). Examples of properties coded as stable are as follows: "They're getting in the freezer and the snowman liked it, 'cause it was nice and cold" and "The sun is hot, right." Examples of properties coded as transitory are as follows: "Are you getting a little warm (referring to body temperature)?" "It's very hot, honey (referring to coffee)."

We found that parents used property-based language for items that are transitory in their temperature (e.g., food) more often than for items that are stable in their temperature (e.g., ice), as shown in **Table 5**. This pattern was observed in all data sets except the Snowman book-reading where parents referred to objects of stable and transitory temperature with similar frequency, which is not surprising given the pictures of fire, ice, snow, and freezers in the book.

TABLE 5 | Percentage of parents' property-based talk that was sub-coded as references to objects with transitory or stable temperatures.

Database	Number of utterances	Transitory	Stable	Significance tests of means
Group data sets				
Snowman study	234	49%	39%	n.s.
Hall	187	84%	1%	$t(39) = 5.41$, $p < 0.001$, $d = 1.20$
Gleason dinner	76	76%	1%	$t(16) = 2.64$, $p < 0.02$, $d = 0.94$
Individual data sets				
Abe	86	84%	6%	
Adam	25	60%	8%	
Eve	40	75%	3%	
Naomi	53	68%	17%	
Sarah	125	83%	8%	

This pattern shows that the ways parents and children talk about heat and temperature likely relate to the activities they are engaged in, which is the focus of the next analysis.

Analysis 3: Contexts for Talk About Heat and Temperature

Parents and children are not likely to engage regularly in formal school-like activities with heat and temperature; we next asked about the everyday situations where conversations about heat and temperature naturally arose in the CHILDES databases.

Main Finding 4

Emphasis on property, process, and substance in discussions of heat and temperature varied depending on the activity context. More specifically, conversations during meal times varied systematically, with more vague process talk while preparing meals and more property talk when eating.

Heat- and temperature-related words occurred most often during mealtimes and in conversations about weather and body temperature, as **Table 6** shows. A descriptive analysis, displayed in **Table 7**, examined which types of talk (property, vague process, and substance) occurred in which types of contexts. Substance-based talk, although infrequent, occurred most often during meal times and conversations about the weather. Property-based talk also occurred most often during meal times and in conversations about the weather, with discussions of body temperature as the third most frequent category. Similarly, vague process talk occurred most often during meal times and in conversations about body temperature, perhaps because these situations relate to children's direct experience with transitions in temperature. Furthermore, the context in which references to the word *heat* occurred most frequently was meal times, and most of these instances reflected the vague process use of the word heat (83%) versus the substance-involved-in-direct-process use (17%) [$\chi^2(1, N = 12) = 5.33$, $p = 0.02$].

Meal times are likely to involve a number of different topics related to temperature, including both process talk (e.g., heating up and cooling down food) and property talk (e.g., discouraging someone from eating something that is too hot). As an exploratory follow-up coding, mealtime contexts were further divided into two categories – meal preparation and eating. The inter-coder reliability on 30% of the mealtime contexts was excellent (Cohen's Kappa = 0.85). Of the 624 utterances coded as mealtime, 236 of them were reliably coded into the meal preparation and eating categories; the remaining 388 utterances could not be further distinguished as talk about eating versus preparing food (e.g., "It's hot tapioca."). Different patterns of property and vague process talk were found in the two contexts. Referring to the temperature-related words as properties occurred in 90% of the eating contexts versus only 33% of the meal preparation contexts. For example, in an

TABLE 6 | Contexts in which heat- and temperature-related talk occurred and the percentage of utterances that were coded as each type of context.

Context	Description	Percentage of Utterances
Meals	Food temperatures. Waiting for food to heat up or cool down.	37%
	Carefully eat hot foods.	
Weather	Temperature of outside/inside. The sun is hot.	17%
Body Temperature	People getting hot/warm/cold.	14%
Dressing	Putting on clothes to stay warm.	6%
	Taking off clothes to get cool.	
	Being too hot or cold wearing some type of clothing	
Touching Objects	Perceiving object temperatures. Getting burned.	4%
Heat Sources	Heat comes from certain objects. People or objects getting too close to heat sources	4%
	Reprimands for touching hot objects.	
Bathing	Taking a bath. Hand washing. Doing dishes.	4%
Swimming	Too cold to go swimming.	1%
Other	Using other meanings of the words.	14%
	Other specific situations such as "warming up" a vehicle.	
	Insufficient information present in the transcripts to determine the context.	

TABLE 7 | Frequency of types of talk observed in each context (excluding "heat").

	Property	Substance	Vague process	Other	Total
Dressing	72	3	34	2	111
Meals	480	10	117	17	624
Weather	249	13	28	3	293
Touching Objects	57	0	3	0	60
Bathing	65	1	5	1	72
Swimming	17	0	0	0	17
Body temp.	190	3	47	3	243
Heat sources	49	5	13	1	68
Other	87	12	14	124	237
Total	1266	47	261	151	1725

eating context, the child said "This is not hot" and the mother replied "No, the lettuce is not hot." In contrast, talking about these words as vague processes occurred in 67% of the meal preparation contexts and in only 10% of the eating contexts [$\chi^2(1, N = 226) = 73.04, p < 0.001$]. For example, in preparing to make candy, the child said, "OK now I think it's finished just for one second" and the mother replied "No it needs to cool a little while longer."

It seems that food preparation may facilitate more discussion of temperature change whereas conversation about eating food may facilitate more discussion of foods having the property of hot or cold. Food preparation activities seem to be a rich context for explicit conversation about the macro-level process of temperature change, beyond the more implicit transitory information that may be conveyed in parents' talk about temperature as properties of objects. The everyday activities in which parents and children engage together seem to afford particular ways of talking about heat and temperature. Interestingly, many of the parents' references to temperature drew upon the children's direct experience with perceiving temperature in their bodies or the ways that temperature affects their interactions with food and other objects.

GENERAL DISCUSSION

The overall goal of this work is to characterize the "data" that children have available in conversations with others so that we can develop informed hypotheses about how these experiences relate to children's developing reasoning about heat and temperature. We sought to describe the nature and the frequency of heat and temperature talk among parents and children. We uncovered a plethora of conversations about heat and temperature in a variety of activities. The naturalistic conversations that we sampled provided evidence of substance-based and process-based talk about *heat*. Both parents and children frequently discussed temperature-related concepts and temperature change. Thus, along with phenomenological experience, children seem to have access to an abundant amount of conversation about heat and temperature.

Part of the inspiration for this research was the intriguing speculation that students' difficulties learning about heat result from their misconstrual of heat as a substance and that this misconstrual could result from hearing heat talked about as a substance in everyday life (Slotta and Chi, 2006). With the current data, we directly tested the latter claim and found that the young children in the sample actually rarely heard the word *heat*, but when parents did use the word they often did imply that heat is a substance or discussed heat as a process involving a substance, which supports Slotta and Chi (2006) claim. Not surprisingly, parents did not use the word *heat* in ways consistent with emergent processes. Despite these findings, there are three striking reasons why the overall picture does not align as clearly with Slotta and Chi's substance-focused prediction as it might appear. First, the vast majority of conversations about heat and temperature involved not the word *heat* but other heat- and temperature-related words such as *hot* and *cold*, which most often functioned to indicate temperature as a property of objects (e.g., "The peas are hot."), and the majority of these references described objects for which temperature is a transient property. Second, most parents and children also had a number of conversations about vague processes related to heat and temperature in certain activities. This leads to the third point: parents' and children's ways of talking about the concepts of heat and temperature varied in meaningful ways across a range of everyday situations, including meal times and getting dressed. We will discuss the implications of these findings below.

In this general discussion, we first discuss the results indicating that parents talk about heat and temperature as a property of objects as well as the findings that these conversations vary by context. Second, we discuss the potential implications of these conversations for children's developing ideas about heat and temperature. Finally, we consider the potential theoretical implications of the findings.

Property, Process, and Context

While parents' mentions of heat were consistent with a substance view, more striking is the low frequency of opportunities to hear heat talked about at all. This supports Lautrey and Mazens (2004) prediction that there may be little support in everyday language for thinking about heat transmission (compared with concepts, although arguably esoteric, such as *resonate* for sound transmission). However, children did have ample opportunity to hear talk about temperature and temperature change. In particular, children in these families were often involved in conversations commenting on *hot* and *cold* as properties of objects. Many of these conversations were about objects that can and do change temperature. Only in the specific context of the Snowman book, with its focus on ice, fire, and snow, were objects with stable temperature talked about as often as objects whose temperature changes often.

Parents' tendency to emphasize temperature as a property of objects varied systematically across contexts. For example, vague processes of temperature change were emphasized in food preparation contexts but properties were emphasized in eating contexts. It is not surprising that parents would emphasize the most relevant aspects of heat/temperature in each situation; for example, when a child is about to put food in her mouth, the end state of the heating process – hot peas – may be more relevant than how the peas became hot. While cooking, however, the process may be of most interest. This variability is consistent with diSessa's notion that a variety of conceptualizations are available and chosen in the moment. Further empirical investigation should explore how children may be making sense of conversations about temperature that systematically vary by context.

While discussions of process were vague and not "scientific" and indeed likely reflect the parents' own non-scientific views, there is evidence that children heard people talk about heat in ways that would direct them to attend to temperature change. In addition, future research and design of science curriculum could take into account the finding that children may be more accustomed to hearing about contact as relevant to heat, compared to other features. Parents in these samples offered

information that heating and cooling are processes that relate to contact with sources of heat and less often to time and proximity. This should be taken into account in later teaching about thermodynamics in school. Understanding that contact is a relevant factor in heat transfer underlies the concept of conduction, which involves transfer of energy through matter. This could perhaps be easier for children than understanding that distance is a relevant factor in the process of heat transfer that underlies the concept of radiation, including the transfer of energy through air or across distances (e.g., the sun's rays radiating to earth). It is true that the kinds of process talk parents used imply "direct processes" of heat transmission on a global level, which Slotta and Chi (2006, p. 263) identify as an inappropriate way to conceptualize the process of heat. While children are not experiencing support for thinking of heat as a molecular emergent process, they are hearing about heat as property and general process and not just as substance.

Implications for Children's Developing Ideas About Heat and Temperature

Because the goal of this work is to characterize children's exposure to language "data" in the form of parents' talk, these findings also result in a set of testable questions for future research regarding how children make use of this "testimony" from adults in developing their own concepts of heat and temperature. Although children infrequently heard parents talk about heat, we do not know for sure whether (and if so, how) these few instances contribute to children's ideas. Children are likely to hear other non-entities (such as time or ideas, see Lakoff and Johnson, 1980) referred to as substances (e.g., "I ran out of time"; "That idea flew out of my mind"). Such talk could lead children to develop misconceptions (see Chi and Hausmann, 2003), however, it is possible that children may instead learn to recognize these figurative uses of language and learn that such language does not always imply substance. This view is consistent with Amin (2009, 2015) argument that children's interpretations of such conceptual metaphors may be an important part of the process of conceptual change whereby embodied schemas are mapped to abstract concepts. It also seems possible that the abundant conversation about temperature and the paucity of conversation about heat may contribute to the conflation of the two concepts observed in prior research (Erickson, 1979; Wiser, 1995) and that when children begin to have formal instruction about "heat" they may try to make sense of the concept in relation to their abundant experience with conceptions of temperature as a property and temperature change.

The prevalence of conversation about *hot* and *cold* as properties is intriguing in light of the early research suggesting that young children (under age 9 years old) view heat and temperature as inherent properties of objects (e.g., Albert, 1978; Erickson, 1980). It is an open question whether these everyday conversations about hot or cold objects would lead children to expect hotness (or heat) and coldness to be inherent properties of objects. Contrary to this view, it seems relevant that we found much of the property-based talk in our samples to describe objects for which hot and cold are transitory properties. Future research is also needed to determine how conversation about hot and cold as properties may encourage substance or general process views of heat. Perhaps a notion of hotness as a property implies that a hot object contains more of the substance called *heat*. However, children's experience with the changing temperature of familiar objects could also suggest to them that *heat* is rather a macro-level process of *heating*. For example, if adults comment on how "hot" the soup is and then a few minutes later tell the child "It's OK to eat it now because it's cool," then one might expect this property talk to encourage a view of heating (or cooling) as a process. This example highlights that it is essential to consider these utterances within the meaningful activity contexts where they unfold. Finally, when parents discussed *hot* and *cold* as changing properties, they rarely elaborated on how things get hotter or colder. Perhaps the language that children hear could still be guiding them to look for explanations of the causes of temperature change (e.g., maybe "hot stuff" left the soup).

The current findings regarding the nature of everyday talk about heat and temperature may have implications for the interpretation of experimental data on children's ideas of heat, hot, and cold. Questions used in earlier interview research, such as "Give me examples of heat" (Albert, 1978), may have been very confusing to young children; their answers may not have demonstrated their full understanding. Future research could explore whether children understand hot and cold differently for objects that are frequently involved in temperature change conversations versus those that are frequently involved in conversations focused on heat as a stable property (e.g., fire).

Parents' and children's tendency to focus on the hotness or the coolness of objects, and on the idea that changing from hot to cold (or cold to hot) is a process, perhaps reflects the everyday goals of ensuring safety when eating foods, deciding what to wear to go outside, and communicating one's body temperature to obtain help in changing it. Using property or direct causal process language to describe the phenomena are perhaps useful ways to communicate about such goals in everyday situations. Although some argue that these kinds of talk are scientifically inaccurate, there seem to be important concepts explicitly and implicitly communicated through the vague process kinds of talk observed in our sample. The notion that everyday talk has overlapping but different goals from scientific talk is reminiscent of Vygotsky's (1987) ideas about spontaneous and scientific concepts. He argued that everyday or "spontaneous" concepts may be incorrect from the perspective of school or "scientific" concepts, and yet children are likely to draw upon these early everyday conceptions as they try to make sense of the concepts that are presented in formal science instruction [see also diSessa (1993b)]. Thus, studies of children's everyday knowledge could be informative for classroom science instruction (e.g., Moschkovich, 2002; Goldman, 2006; National Research Council, 2007). Moreover, scientists' explanations of everyday phenomena regarding heat may have more in common with students' explanations than one might think (Lewis and Linn, 1994), and these intuitive concepts should be taken seriously when thinking about formal science instruction (Hammer and Van Zee, 2006).

Implications for Theory

We do not believe that these data support one theory over another. They are rather a systematic look at the actual linguistic information regarding heat and temperature available to young children, which various theories speculate about. The data are important in that they may indicate productive points of overlap between competing theories, and we suggest that a sociocultural perspective can help us attend to such overlap. In this section, we first speculate about how the data bear on the three main theories discussed earlier and, in conclusion, suggest points of overlap in attempts to move beyond pitting one theory against another.

Ontological Knowledge

While we found that not all talk about heat was substance-based, our findings generally support Slotta and Chi's (2006) claim that everyday talk about heat does not overlap much with their conception of "scientific" talk about heat. The data clearly do not support an argument that parents teach children an emergent process way of thinking. On the other hand, the finding that everyday talk supports not just substance but property and vague process views of heat as well, opens up further questions regarding how parents' talk may influence children's developing everyday views of temperature and what the links as well as disconnects are between these everyday views and the later views learned in school. When considering development of reasoning about heat, an ontological view might further investigate the significance of the varied ways that children encounter different aspects of heat and temperature. The data show that children hear a lot of talk about temperature as properties of objects that can and do change in temperature frequently. How might children think differently about liquids versus solids, food versus non-edible objects, or the fluctuations in air temperature in a single day versus across a year with distinct seasons? How do ontological frames apply to the various situations that children encounter?

Theory Theory

Scholars who favor a theory theory approach are focused on the coherence of children's thinking, regardless of whether it is "accurate" or similar to adult thinking (Carey, 1985). If children move from a conception of heat as a substance to heat as a process, then theory theory approaches would suggest that this is embedded in a larger intuitive theory change. If such a global theory change occurs, then our data on everyday ways of talking about heat would need to be taken into account as part of the picture of children's changing conceptualizations of heat and temperature. How might children's ideas be organized when thinking about different kinds of temperature-related phenomena? Might they have seemingly separate coherent theories about how air temperature changes versus how liquids heat up and cool down?

Another focus of theory theory is children's sensitivity to the causal basis for events in the natural world (Gopnik, 2000). In some sense, the focus on causality in naïve theories could be seen as an additional barrier to learning the ontological category of emergent process because the type of causal thinking that is intuitive to people resembles direct causal processes. Children's task in forming coherent theories of heat and temperature would involve integrating the various conceptualizations of temperature that they hear across many different contexts with their own experiences to construct a coherent view of these phenomena, including a mechanism for how temperature change works. Our data suggest that children hear little explicit discussion of what this causal mechanism might be and how it works. The fact that parents seem to be using multiple conceptual structures in the language that they use to describe heat, temperature, heating, and cooling likely makes the task of forming a coherent theory more difficult for children. This is perhaps exacerbated by the fact that the parents' own theories of heat and temperature are based on intuitive views that are also likely inconsistent with the scientific view.

"Knowledge in Pieces"

From a "knowledge-in-pieces" view, the variation and the contradictions in the information available to children in family conversations might not be problematic for children's developing understanding of heat and temperature because this is the way people are argued to think in general. In a recent study, for example, diSessa (2017) argues that the knowledge-in-pieces approach best explains the conceptual change regarding thermal equilibration observed in a high school classroom. By this approach, the language children hear in everyday conversations with their parents may offer children access to "pieces" of knowledge that they may later employ in different ways in different situations and would be consistent with studies showing that children's answers to heat- and temperature-related questions vary depending upon the context and the method of assessment (Jones et al., 2000; Clark, 2006). In this view, the explanations used by parents and children in any given context are meaningfully related to and inseparable from those contexts because the explanations are also dependent upon the features of the situation, the goal of the activity, and the experience-based knowledge that each context can cue.

Conclusion

Rather than attempting to decide among these dynamically changing theories, it may be more productive to find areas of agreement as we move forward to find a new framework. Including a sociocultural perspective may be useful in beginning to integrate the existing perspectives. A sociocultural view assumes that learning is situated and cannot be separated from the social and the cultural context (Rogoff, 1990; Lave and Wenger, 1991; Rogoff et al., 2018). Depending upon the particular context in which they are thinking, physicists and children may conceptualize heat in different ways (e.g., substance, direct causal process, emergent process) to accomplish the goals of the activity. Another point made by sociocultural theorists is that there is no pure way to measure underlying conceptions, and researchers should be wary about assuming that any particular assessment gives us the "true" measure of a person's concept (Lave, 1997; Rogoff, 1997). Assessment can be viewed instead as documenting the ways that children participate in everyday activities and how such participation changes over time. Each experiment or naturalistic activity offers one glimpse of what children

do, which may be more informative than trying to extract what concepts children *have*. From this stance, how children participate in everyday conversations is a necessary component of understanding the ways that their knowledge about heat and temperature is developing: *how* children's knowledge develops is by participating in activities (Wertsch, 1979). A final relevant point from sociocultural theory is that attending to different aspects of sociocultural activity can allow for insight into varied ways of understanding learning and development (Rogoff, 1997). Applying this idea, a knowledge-in-pieces lens can help us understand how individuals are drawing upon a vast set of resources to reason in any given moment. These can include various modes of thinking, the language used to communicate ideas, and prior experiences with the phenomena itself and with communicating *about* the phenomena. A theory theory lens on the same behavior can help us understand how children's thinking can appear to be organized consistently (or not) across contexts and how it can appear that consistency in reasoning changes over time. Rather than attributing the consistency to stable, coherent theories that persist across contexts, a sociocultural view would suggest that the consistency is related to the social contexts and kinds of situations that children reason in. The child-participating-in-context may be the consistency that is often attributed solely to the child. So we can then ask how children's participation in reasoning activities changes over time instead of how children's internal theories change.

Our goal is not to argue for a unitary direct causal path between parents' speech and children's reasoning about heat and temperature. We rather assume that children are active in constructing their understanding of the phenomena. We aimed to map out one crucial aspect of the terrain – the available resources in the everyday talk that children engage in and which they can draw upon in developing understandings about heat and temperature. The data that we sampled indicated that there is systematicity in the language that the children heard. The next steps involve investigating how children integrate the language that they hear with their physical and conceptual experiences involving heat and temperature.

Taking participation in social activity as a potential *mechanism* of cognitive development and learning, theoretical views should more fully take into account the everyday activities and uses of the terms *heat* and *temperature* in conversations and how children's participation in such activities changes over time, especially when they encounter formal instruction. Clearly the present data support the notion that even very young children view heat and temperature as a topic with which they have much familiarity and their intuitive notions have been helpful to them as they navigate their physical world. We would argue that rather than focusing on replacing children's experience-based knowledge (or "misconceptions") with "correct" scientific knowledge, it may be more productive to consider the approach taken by diSessa and colleagues (Smith et al., 1993) as well as Wiser and Amin (2001) that there are different, valid ways of thinking about phenomena such as heat and temperature (scientific and everyday) and that learning can build on the intuitive notions. Even within theory theory approaches, the idea that inconsistent theories can co-exist in people's minds has gained acceptance (Legare et al., 2012; Shtulman and Lombrozo, 2016). In support of this approach, educational researchers have theorized that meaningful learning can take place in "hybrid spaces" or "third spaces," which are neither everyday nor school, where children are encouraged to use their everyday ideas to help them start to learn scientific concepts or dominant practices (Gutiérrez et al., 1999; Rosebery et al., 2005). For example, settings such as after-school programs and museums have been argued to be places where children may begin to learn to bridge their everyday and scientific views of concepts such as heat and temperature. Our study has demonstrated the importance of investigating young children's everyday linguistic experiences in efforts to understand how such integrations of everyday and scientific thinking occur.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of California Santa Cruz Institutional Review Board. Written informed consent to participate in this study was provided by the participants' parents or legal guardians.

AUTHOR CONTRIBUTIONS

ML developed the conceptual motivation for the work, completed coding and statistical analysis for both the CHILDES study and the book-reading study, and wrote initial drafts of the manuscript. MC collaborated with two other researchers in designing the book-reading study and supervised data collection for that study, collaborated on the design of coding and analysis for both studies, and collaborated on writing and revising the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

ML's work was supported by the NSF grant ESI0119787 as part of the Center for Informal Learning and Schools, and by University of California, Santa Cruz. The parent-child book-reading study was originally supported by the NIH grant HD26228 to MC and by the Institute for Research on Learning.

REFERENCES

Abimbola, I. (1988). The problem of terminology in the study of student conceptions in science. *Sci. Educ.* 7, 175–184. doi: 10.1002/sce.3730720206

Albert, E. (1978). Development of the concept of heat in children. *Sci. Educ.* 62, 389–399. doi: 10.1002/sce.3730620316

Amin, T. G. (2009). Conceptual metaphor meets conceptual change. *Hum. Dev.* 52, 165–197. doi: 10.1159/000213891

Amin, T. G. (2015). Conceptual metaphor and the study of conceptual change: research synthesis and future directions. *Intern. J. Sci. Educ.* 37, 966–991. doi: 10.1080/09500693.2015.1025313

Bakeman, R., and Gottman, J. M. (1986). *Observing Interaction: An Introduction To Sequential Analysis*. New York, NY: Cambridge University Press.

Briggs, R. (1978). *The Snowman*. New York, NY: Random House, Inc.

Brookes, D. T., and Etkina, E. (2015). The importance of language in students' reasoning about heat in thermodynamic processes. *Intern. J. Sci. Educ.* 37, 759–779. doi: 10.1080/09500693.2015.1025246

Brown, R. (1973). *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Bruner, J. (1983). *Child's taLk: Learning to Use Language*. New York, NY: W. W. Norton & Company.

Callanan, M., and Jipson, J. L. (2001). "Explanatory conversations and young children's developing scientific literacy," in *Designing for Science: Implications From Everyday, Classroom, And Professional Settings*, eds K. Crowley, C. D. Schunn, and T. Okada (Mahwah, NJ: Erlbaum), 21–49.

Callanan, M. A., and Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cogn. Dev.* 7, 213–233. doi: 10.1016/0885-2014(92)90012-g

Callanan, M. A., Shrager, J., and Moore, J. L. (1995). Parent-child collaborative explanations: Methods of identification and analysis. *J. Learn. Sci.* 4, 105–129. doi: 10.1207/s15327809jls0401_3

Carey, S. (1985). *Conceptual Change In Childhood*. Cambridge, MA: MIT Press.

Carey, S. (1991). "Knowledge acquisition: Enrichment or conceptual change?," in *The Epigenesis of Mind: Essays on Biology And Cognition*, eds S. Carey and R. Gelman (Hillsdale, NJ: Lawrence Erlbaum), 257–291.

Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: why some misconceptions are robust. *J. Learn. Sci.* 14, 161–199. doi: 10.1207/s15327809jls1402_1

Chi, M. T. H., and Hausmann, R. G. M. (2003). "Do radical discoveries require ontological shifts?," in *The International Handbook On Innovation*, ed. L. V. Shavinina (New York, NY: Elsevier Science), 430–444. doi: 10.1016/b978-008044198-6/50030-9

Clark, D., and Jorde, D. (2004). Helping students revise disruptive experientially supported ideas about thermodynamics: computer visualizations and tactile models. *J. Res. Sci. Teach.* 41, 1–23. doi: 10.1002/tea.10097

Clark, D. B. (2006). Longitudinal conceptual change in students' understanding of thermal equilibrium: An examination of the process of conceptual restructuring. *Cogn. Instr.* 24, 467–563. doi: 10.1207/s1532690xci2404_3

Clough, E., and Driver, R. (1985). Secondary students' conceptions of the conduction of heat: bringing together scientific and personal views. *Phys. Educ.* 20, 176–182. doi: 10.1088/0031-9120/20/4/309

Crowley, K., Callanan, M. A., Jipson, J. L., Galco, J., Topping, K., and Shrager, J. (2001). Shared scientific thinking in everyday parent-child activity. *Sci. Educ.* 85, 712–732. doi: 10.1002/sce.1035

Crowley, K., and Jacobs, M. (2002). "Building islands of expertise in everyday family activity," in *Learning Conversations In Museums*, eds G. Leinhardt, K. Crowley, and K. Knutson (Mahwah, NJ: Lawrence Erlbaum Associates), 333–356.

diSessa, A. (1988). "Knowledge in pieces," in *Constructivism in the Computer Age*, eds G. Forman and P. B. Pufall (Hillsdale, NJ: Lawrence Erlbaum), 49–70.

diSessa, A. (1993a). Ontologies in pieces: response to Chi and Slotta. *Cogn. Instruct.* 10, 261–280.

diSessa, A. (1993b). Toward an epistemology of physics. *Cogn. Instruct.* 10, 105–225. doi: 10.1207/s1532690xci1002%263_2

diSessa, A. A. (1996). "What do "just plain folk" know about physics?," in *The Handbook Of Education And Human Development: New Models Of Learning, Teaching And Schooling*, eds D. R. Olson and N. Torrance (Malden, MA: Blackwell Publishers), 709–730.

diSessa, A. A. (2000). *Changing Minds: Computers, Learning, And Literacy*. Cambridge, MA: MIT Press.

diSessa, A. A. (2017). Conceptual change in a microcosm: comparative analysis of a learning event. *Hum. Dev.* 60, 1–37. doi: 10.1159/000469693

Erickson, G., and Tiberghien, A. (1985). "Heat and temperature," in *Children's Ideas In Science*, eds R. Driver, E. Guesne, and A. Tiberghien (Philadelphia: Open University Press), 52–84.

Erickson, G. L. (1979). Children's conceptions of heat and temperature. *Sci. Educ.* 63, 221–230. doi: 10.1002/sce.3730630210

Erickson, G. L. (1980). Children's viewpoints of heat: a second look. *Sci. Educ.* 64, 323–336. doi: 10.1002/sce.3730640307

Gelman, S. A. (2009). Learning from others: children's construction of concepts. *Annu. Rev. Psychol.* 60, 115–140. doi: 10.1146/annurev.psych.59.103006.093659

Gleason, J. B., Perlmann, R. Y., and Greif, E. B. (1984). What's the magic word? Learning language through routines. *Discour. Process.* 6, 493–502. doi: 10.1080/01638538409544603

Goldman, S. (2006). "A new angle on families: connecting the mathematics of life with school mathematics," in *Learning in Places: The Informal Education Reader*, eds Z. Bekerman, N. C. Burbules, and D. S. Keller (New York, NY: Peter Lang Publishing), 55–76.

Gopnik, A. (2000). "Explanation as orgasm and the drive for causal knowledge: the function, evolution, and phenomenology of the theory formation system," in *Explanation and Cognition*, eds F. C. Keil and R. A. Wilson (Cambridge, MA: MIT Press), 299–323.

Gopnik, A., and Meltzoff, A. N. (2002). *Words, Thoughts, And Theories*. Cambridge, MA: Massachusetts Institute of Technology Press.

Gupta, A., Hammer, D., and Redish, E. F. (2010). The case for dynamic models of learners' ontologies in physics. *J. Learn. Sci.* 19, 285–321. doi: 10.1080/10508406.2010.491751

Gutiérrez, K. D., Baquedano-López, P., and Tejada, C. (1999). Rethinking diversity: hybridity and hybrid language practices in the third space. *Mind Cult. Activ.* 6, 286–303. doi: 10.1080/10749039909524733

Hall, W. S., and Tirre, W. C. (1979). *The Communicative Environment Of Young Children: Social Class, Ethnic And Situational Differences*. Champaign, IL: University of Illinois.

Hammer, D. M. (2000). Student resources for learning introductory physics. *Am. J. Phys. Edu. Res. Suppl.* 68, S52–S59.

Hammer, D. M., and Van Zee, E. (2006). *Seeing the Science In Children's Thinking*. New York, NY: Heinemann.

Harris, P. L. (2002). "What do children learn from testimony?," in *The Cognitive Basis Of Science*, eds P. Carruthers, M. Siegal, and S. Stich (Cambridge: Cambridge University Press).

Harris, P. L., and Koenig, M. A. (2006). Trust in testimony: how children learn about science and religion. *Child Dev.* 77, 505–524. doi: 10.1111/j.1467-8624.2006.00886.x

Hobson, A. (1995). Heat is not a noun. *Phys. Teach.* 33, 325–326.

Jeppson, F., Frejd, J., and Lundmark, F. (2017). "Wow, it turned out red! First, a little yellow, and then red!" 1st graders' work with an infrared camera. *J. Res. Childh. Educ.* 31, 581–596. doi: 10.1080/02568543.2017.1347589

Jones, M. G., Carter, G., and Rua, M. J. (2000). Exploring the development of conceptual ecologies: communities of concepts related to convection and heat. *J. Res. Sci. Teach.* 37, 139–159. doi: 10.1002/(sici)1098-2736(200002)37:2<139::aid-tea4>3.0.co;2-1

Keil, F. C. (1979). *Semantic and Conceptual Development*. Cambridge, MA: Harvard University Press.

Keil, F. C. (1983). On the emergence of semantic and conceptual distinctions. *J. Exper. Psychol. Gen.* 112, 357–385. doi: 10.1037/0096-3445.112.3.357

Kuczaj, S. (1976). *-ing, -s and -ed: A Study Of The Acquisition Of Certain Verb Inflections*. Doctoral dissertation, University of Minnesota, Minnesota.

Lakoff, G., and Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cogn. Sci.* 4, 195–208. doi: 10.1207/s15516709cog0402_4

Lautrey, J., and Mazens, K. (2004). Is children's naïve knowledge consistent? A comparison of the concepts of sound and heat. *Learn. Instruct.* 14, 399–423. doi: 10.1016/j.learninstruc.2004.06.011

Lave, J. (1997). "What's special about experiments as contexts for thinking," in *Mind, Culture, and Activity: Seminal Papers From The Laboratory of Comparative Human Cognition*, eds M. Cole, Y.

Engström, and O. Vasquez (New York, NY: Cambridge University Press), 57–69.

Lave, J., and Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Legare, C., Evans, E. M., Rosengren, K., and Harris, P. (2012). The coexistence of natural and supernatural explanations across cultures and development. *Child Dev.* 83, 779–783.

Lewis, E. L., and Linn, M. C. (1994). Heat energy and temperature concepts of adolescents, adults, and experts: implications for curricular improvements. *J. Res. Sci. Teach.* 31, 657–677. doi: 10.1002/tea.3660310607

MacWhinney, B. (1995). *The CHILDES Project: Tools for Analyzing Talk*, 2nd Edn, Hillsdale, NJ: Lawrence Erlbaum.

Mazens, K., and Lautrey, J. (2003). Conceptual change in physics: children's naive representations of sound. *Cogn. Dev.* 18, 159–176. doi: 10.1016/s0885-2014(03)00018-2

McCloskey, M. (1983). "Naïve theories of motion," in *Mental Models*, eds D. Gentner and A. Stevens (Hillsdale, NJ: Lawrence Erlbaum), 289–324.

Moschkovich, J. (2002). A situated and sociocultural perspective on bilingual mathematics learners. *Math. Think. Learn.* 4, 189–212. doi: 10.1207/s15327833mtl04023_5

National Research Council (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8*. Washington, DC: The National Academies Press. doi: 10.17226/11625

Piaget, J. (1974). *Understanding Causality*. New York, NY: W. W. Norton.

Reif, F. (1965). *Fundamentals of Statistical And Thermal Physics*. New York, NY: McGraw-Hill.

Reiner, M., Slotta, J. D., Chi, M. T. H., and Resnick, L. B. (2000). Naïve physics reasoning: a commitment to substance-based conceptions. *Cogn. Instruct.* 18, 1–34. doi: 10.1207/s1532690xci1801_01

Rogoff, B. (1990). *Apprenticeship In Thinking: Cognitive Development In Social Context*. New York, NY: Oxford University Press.

Rogoff, B. (1997). "Evaluating development in the process of participation: theory, methods, and practice building on each other," in *Change and Development*, eds E. Amsel and A. Renninger (Hillsdale, NJ: Erlbaum), 265–285.

Rogoff, B., Dahl, A., and Callanan, M. (2018). The importance of understanding children's lived experience. *Dev. Sci.* 50, 5–15. doi: 10.1016/j.dr.2018.05.006

Rosebery, A. S., Ogonowski, M., DiSchino, M., and Warren, B. (2010). "The coat traps all your body heat": heterogeneity as fundamental to learning. *J. Learn. Sci.* 19, 322–357. doi: 10.1080/10508406.2010.491752

Rosebery, A. S., Warren, B., Ballenger, C., and Ogonowski, M. (2005). "The generative potential of students' everyday knowledge in learning science," in *Understanding Mathematics And Science Matters: Studies In Mathematical Thinking And Learning*, eds T. A. Romberg, T. P. Carpenter, and F. Dremock (Mahwah, NJ: Lawrence Erlbaum), 55–79.

Sachs, J. (1983). "Talking about the there and then: the emergence of displaced reference in parent-child discourse," in *Children's Language*, Vol. 4, ed. K. E. Nelson (Hillsdale, NJ: Lawrence Erlbaum Associates).

Shtulman, A. (2017). *Scienceblind: Why Our Intuitive Theories About The World Are So Often Wrong*. New York, NY: Basic Books.

Shtulman, A., and Lombrozo, T. (2016). "Bundles of contradictions: a coexistence view of conceptual change," in *Core Knowledge And Conceptual Change*, eds D. Barner and A. Baron (Oxford, UK: Oxford University Press).

Siegler, R. S. (2007). Cognitive variability. *Dev. Sci.* 10, 104–109.

Slotta, J. D., and Chi, M. T. H. (2006). Helping students understand challenging topics in science through ontology training. *Cogn. Instruct.* 24, 261–289. doi: 10.1207/s1532690xci2402_3

Slotta, J. D., Chi, M. T. H., and Joram, E. (1995). Assessing students' misclassifications of physics concepts: an ontological basis for conceptual change. *Cogn. Instruct.* 13, 373–400. doi: 10.1207/s1532690xci1303_2

Smith, J. P., diSessa, A. A., and Roschelle, J. (1993). Misconceptions reconceived: a constructivist analysis of knowledge in transition. *J. Learn. Sci.* 3, 115–163. doi: 10.1207/s15327809jls0302_1

Snow, C. E., and Kurland, B. F. (1996). "Sticking to the point: talk about magnets as a context for engaging in scientific discourse," in *Discourse, Learning, And Schooling*, ed. D. Hicks (New York, NY: Cambridge University Press), 189–220. doi: 10.1017/cbo9780511720390.007

ter Haar, D., and Wergeland, H. (1960). *Elements of Thermodynamics*. Reading, MA: Addison-Wesley.

Tizard, B., and Hughes, M. (1984). *Young Children Learning*. Cambridge, MA: Harvard University Press.

Vosniadou, S. (2009). 'Conceptual metaphor meets conceptual change': yes to embodiment, no to fragmentation. *Hum. Dev.* 52, 198–204. doi: 10.1159/000213892

Vosniadou, S., and Brewer, W. F. (1992). Mental models of the earth: a study of conceptual change in childhood. *Cogn. Psychol.* 24, 535–585. doi: 10.1016/0010-0285(92)90018-w

Vygotsky, L. S. (1978). *Mind in Society: The Development Of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1987). "The development of scientific concepts in childhood," in *The Collected Works of L.S. Vygotsky*, Vol. 1, eds R. W. Rieber, A. S. Carton, and N. Minick (New York, NY: Plenum), 167–241.

Wellman, H. M., and Gelman, S. A. (1998). "Knowledge acquisition in foundational domains," in *Handbook of Child Psychology: Volume 2: Cognition, Perception, and Language*, ed. W. Damon (Hoboken, NJ: John Wiley & Sons), 523–573.

Wertsch, J. V. (1979). From social interaction to higher psychological processes. A clarification and application of Vygotsky's theory. *Hum. Dev.* 22, 1–22. doi: 10.1159/000272425

Wilkening, F., and Huber, S. (2004). "Children's intuitive physics," in *Blackwell Handbook Of Childhood Cognitive Development*, ed. U. Goswami (Oxford, UK: Blackwell), 349–370.

Wiser, M. (1995). "Use of history of science to understand and remedy students' misconceptions about heat and temperature," in *Software Goes To School: Teaching For Understanding With New Technologies*, eds D. N. Perkins, J. L. Schwartz, M. M. West, and M. S. Wiske (New York, NY: Oxford University Press), 23–38. doi: 10.1093/acprof:oso/9780195115772.003.0002

Wiser, M., and Amin, T. (2001). Is heat hot? Inducing conceptual change by integrating everyday and scientific perspectives on thermal phenomena. *Learn. Instruct.* 11, 331–355. doi: 10.1016/s0959-4752(00)00036-0

Wiser, M., and Carey, S. (1983). "When heat and temperature were one," in *Mental Models*, eds D. Gentner and A. L. Stevens (Hillsdale, NJ: Lawrence Erlbaum), 267–297.

Zemansky, M. W. (1970). The use and misuse of the word "heat" in physics teaching. *Phys. Teach.* 8, 295–300. doi: 10.1119/1.2351512

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Luce and Callanan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 25 August 2020
doi: 10.3389/fpsyg.2020.02210

Of Blickets, Butterflies, and Baby Dinosaurs: Children's Diagnostic Reasoning Across Domains

Deena Skolnick Weisberg[1], Elysia Choi[2] and David M. Sobel[3]*

[1] Department of Psychological and Brain Sciences, Villanova University, Villanova, PA, United States, [2] Department of Applied Psychology, New York University, New York, NY, United States, [3] Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, United States

The three studies presented here examine children's ability to make diagnostic inferences about an interactive causal structure across different domains. Previous work has shown that children's abilities to make diagnostic inferences about a physical system develops between the ages of 5 and 8. Experiments 1 (N = 242) and 2 (N = 112) replicate this work with 4- to 10-year-olds and demonstrate that this developmental trajectory is preserved when children reason about a closely matched biological system. Unlike Experiments 1 and 2, Experiment 3 (N = 110) demonstrates that children struggle to make similar inferences when presented with a parallel task about category membership in biology. These results suggest that children might have the basic capacity for diagnostic inference at relatively early ages, but that the content of the inference task might interfere with their ability to demonstrate such capacities.

Keywords: diagnostic inference, scientific reasoning, early elementary school, causal reasoning, contextualization

OPEN ACCESS

Edited by:
Amanda C. Brandone, Lehigh University, United States

Reviewed by:
Paul Muentener, Tufts University, United States
Igor Bascandziev, Harvard University, United States

****Correspondence:***
Deena Skolnick Weisberg
deena.weisberg@villanova.edu

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *29 January 2020*
Accepted: *06 August 2020*
Published: *25 August 2020*

Citation:
Weisberg DS, Choi E and Sobel DM (2020) Of Blickets, Butterflies, and Baby Dinosaurs: Children's Diagnostic Reasoning Across Domains. Front. Psychol. 11:2210. doi: 10.3389/fpsyg.2020.02210

INTRODUCTION

Diagnostic reasoning — inferring causes from systematic observations of patterns of data — is a hallmark of scientific thinking. It involves reasoning backwards, often from observed effects to potential causes. If our car doesn't start, we wonder if there is fuel in the tank. If we smell gas in our apartment, we check the stove to find out whether there is a leak. If a soufflé doesn't rise, we wonder whether we added enough cream of tartar. As adults, we easily engage in this kind of diagnostic reasoning (e.g., Thomas et al., 2008; Fernbach et al., 2010, 2011), which can be construed as a form of causal reasoning (Gopnik et al., 2001).

Research on children's causal reasoning shows that preschoolers, and in some cases infants, can diagnose causal structures, draw inferences about the nature of causal systems from observed data, and even explore in systematic ways that both reflect appropriate causal inferences and generate new pieces of knowledge (Gopnik et al., 2004; Koerber et al., 2005; Sobel and Kirkham, 2006; Schulz and Bonawitz, 2007; Cook et al., 2011). Although this work seems to suggest that diagnostic reasoning is present early, in all of these studies, children or babies observe the efficacy of all or all but one potential cause in a system. Children's thinking shows a more pronounced developmental trajectory when asked to reason about systems with multiple uncertain causes (Beck et al., 2006; Fernbach et al., 2012; Erb and Sobel, 2014; Sobel et al., 2017).

As an example, Sobel et al. (2017) showed 5- to 8-year-olds a pattern of data generated by an interactive causal model where the efficacy of some causes was left unspecified. Specifically, children

in this study were asked to reason about a blicket detector machine (Gopnik and Sobel, 2000): a box that lights up and plays music when objects like blocks are placed on it, giving the illusion that the blocks make the box light up. Children were shown four possible causes (different blocks) that could cause one of three different kinds of activation on the machine (off, red, or green with music). In this system, two of the blocks are jointly necessary for the target effect (green with music), but these two blocks have a different effect (red) when they are not used together. This kind of reasoning problem parallels the inferences one would have to make in scientific reasoning; indeed, the causal structure of this blicket detector system was based on a model presented to older children in a test of scientific thinking called *Earthquake Forecaster* (Kuhn et al., 2009). Sobel et al. (2017) found clear developmental differences between 5 and 8 years in children's reasoning about this system. The 5- and 6-year-olds in this study responded no differently from chance, while the 7- and 8-year-olds were successful at diagnosing the causal structure.

Engaging in diagnostic reasoning in tasks like this one requires a set of cognitive and metacognitive capacities, which have been well-documented in both the literature on cognitive development and on children's developing scientific reasoning (Klahr, 2002) (see e.g., Zimmerman, 2000; Kuhn, 2007). But in addition to these general cognitive abilities, in order to diagnose potential causes, one may also need domain-specific knowledge in order to understand which events are potential causes of observed data. For example, to diagnose why our soufflé didn't rise, we must know that not adding enough cream of tartar will cause a flat soufflé. We must contrast that possibility with the possibility that we curdled the eggs, overwhipped them, or failed to bring them to room temperature prior to whipping. To do this, we must know that these are also potential causes of flat soufflés (among numerous others). That is, we use our prior knowledge to form a hypothesis space in which we reason diagnostically (Sobel et al., 2004; Goodman et al., 2011; Griffiths et al., 2011; Ullman et al., 2012).

Given this, results from studies on infants and young children suggest that they have the cognitive capacity to engage in diagnostic reasoning only under optimal conditions, where they not only are presented with data sets that lack uncertainty but also where they are not required to bring more than the most general prior knowledge to bear on solving the problem. Indeed, the blicket detector paradigm, which is used in much of the prior work on young children's causal reasoning abilities, asks children to reason about physical causal relations that they can see and articulate (Sobel and Munro, 2009a). This paradigm purposely aims to minimize and control the amount of prior knowledge that is necessary for children to successfully make diagnostic inferences (see, e.g., Schulz and Sommerville, 2006; Schulz and Bonawitz, 2007; Schulz et al., 2007; Sobel and Sommerville, 2009b; Bonawitz et al., 2011, for other examples).

It is thus an open question whether children are still successful at diagnostic reasoning when the context of the reasoning task is changed in a way that might affect how they construct the potential hypothesis space – specifically, by adding real-world information that might interact with the general reasoning abilities that children bring to bear in thinking about the task. Across three experiments, we investigated whether a task's surface features – specifically whether those surface features seem to require the child to enlist their domain-specific knowledge – would affect children's diagnostic inferences. Importantly, the structure of the reasoning problem presented in all of the current experiments was identical; no domain-specific biological knowledge was required to solve any of our tasks. However, because some versions of the task couched it in terms of biological systems, they may have encouraged children to bring such knowledge to bear on the problem. Our general goal was to investigate whether this would affect children's reasoning abilities, either positively or negatively.

To do so, we first replicated the Sobel et al. (2017) procedure using a blicket detector with a wider age range and larger sample size. We also presented children with an analogous diagnostic reasoning problem with different surface features (butterflies and flowers). We compared these two tasks in both a between subject (Experiment 1) and a within-subject (Experiment 2) design. Our goal in these first two experiments was to see whether children would make the same kind of diagnostic inference in a domain other than physical causality and, if so, how those inferences compared. Results from these two experiments can thus illustrate whether minimal changes to the surface features of a diagnostic inference task might affect children's reasoning abilities.

Experiment 3 moved beyond these tasks in several ways. We presented children with a diagnostic inference task that was about category membership, instead of about causal relations, but that had the same underlying structure as the systems in Experiments 1 and 2. Specifically, Experiment 3 asked children to make diagnostic inferences about category membership of a set of dinosaurs. Although the interactive structure was identical to the reasoning problems in Experiments 1 and 2, numerous aspects of the procedure differed. The task introduced potentially novel vocabulary (e.g., new dinosaur names) and embedded the problem in a more traditional scientific framework. Moreover, children had to track different features as potentially relevant, instead of different levels of a single feature (color in Experiments 1 and 2). All of these changes added to the information processing capacities necessary to make a diagnostic inference.

As such, Experiment 3 was deliberately designed to differ from Experiments 1 and 2 in a variety of ways. Our main goal in this experiment was not to investigate the impact of individual features or combinations of features on children's performance in a diagnostic reasoning task. Rather, we aimed to translate our causal system into a form that more closely resembled scientific thinking problems that have been previously used to assess children's reasoning abilities (e.g., Kuhn and Dean, 2005; see also Kuhn et al., 2009). Because children might encounter problems like this one in a classroom or a genuine science lab, this dinosaur version of the task serves as an important bridge between work with the blicket detector systems typically used in cognitive developmental psychology and work with the more realistic systems typically used in education science.

Experiment 3 can thus help us to determine whether children's diagnostic reasoning capacities are robust and domain-general, in which case they should show a similar developmental trajectory to Experiments 1 and 2, or whether more realistic

presentations of a diagnostic reasoning problem can affect their performance. That is, although the current design does not allow us to determine which particular demand characteristics may influence children's reasoning, it does allow us to examine whether children possess domain-general diagnostic reasoning abilities or whether those abilities are limited by information processing demands or domain-specific knowledge. No prior work, to our knowledge, has presented a causal system previously studied with blicket detectors (as in developmental psychology) within a realistic scientific framework (as in education science), allowing this experiment to begin to build a bridge between these two literatures.

FIGURE 1 | Stimuli for the blicket **(top)** and butterfly **(bottom)** conditions. The bottom left picture shows the back side of the butterfly apparatus. The experimenter would lift the correct color of butterflies upwards and the butterflies would appear to the participant as in the bottom middle picture. Because the green butterflies also were accompanied by music, a speaker that plays music when pressed is attached to the left side of handle.

EXPERIMENT 1

In Experiment 1, we replicated the procedure used by Sobel et al. (2017) and extended it to a second diagnostic reasoning problem with the same causal structure. This second problem – about whether butterflies were attracted to different colors of flowers – was designed to be as similar to the blicket detector version of the procedure as possible, but instead of asking about the physical relation between objects and a machine, it presented a task about biological mechanisms. Of interest is (a) whether we replicate the finding demonstrated by Sobel et al. (2017) that children begin to succeed at diagnostic inference tasks of this nature around age 7, and (b) whether children show similar development for a measure with different surface features, but the same underlying causal structure.

Method

Participants

The final sample consisted of 242 children between the ages of 4 and 10 (118 boys, 124 girls, M_{age} = 83.09, SD = 21.54). This sample size was chosen to parallel the sample size of Sobel et al. (2017, Study 3). Ten additional children were tested, but not included in the final sample. Nine were excluded because of experimental error and one was not fluent in English. Children were tested at a local science museum and several preschools in the Philadelphia area. The racial breakdown of the sample was as follows: 153 families were White/Caucasian, 25 were Black/African American, 22 were of Asian descent, 1 was of Native American descent, 13 were of mixed descent and 28 did not report this information. The ethnic breakdown of the sample was as follows: 18 families reported as Hispanic, 133 reported as not Hispanic, and 91 families did not report this information.

Materials

We used a blicket detector, which is a remote-controlled, battery-powered rectangular box (19.5 cm × 15 cm × 7.8 cm). The box was black with a white pressure-sensitive plate on top (see **Figure 1**, top left panel). Pressing this plate would trigger a set of LED lights under it, which would make the machine turn on different colored lights. The machine could also play music from an internal speaker when activated. A second experimenter, who sat behind the first experimenter running the study, controlled the blicket using a hidden remote. This remote was used to first activate the machine (so that placing objects on the pressure-sensitive plate would make it turn on) and then was used to change the color of the activation (to red or to green, depending on the trial; see below).

This condition also used four small square canvases stretched over thick wooden frames (which we called "blocks") painted white, blue, orange, and black (5 cm × 5 cm; see **Figure 1**, top right panel). There was also a clear cylindrical container in which the blocks could fit (height 8 cm, radius of base 4.25 cm). We also created a set of 7 laminated photographs of different combinations of the colored blocks (9.5 cm × 7.25 cm). These were used as visual reminders of each step of the procedure (4 photographs) and as possible responses to our test question (3 photographs; see Procedure section).

We also used four silk flowers in the colors white, black, orange, and blue. The flowers were all of the same type and size (12 cm radius) and were purchased from a local craft store. We glued these flowers onto wooden dowels, which were painted green in order to look like flower stems (20 cm). We constructed a flower pot in which to "plant" the flowers, using a rectangular block of foam (31 cm × 11 cm × 11 cm), which we painted yellow and brown (see **Figure 1**, bottom right panel). We punched holes in the top of the foam in order to "plant" the flowers.

We also used 6 plastic butterflies, 3 green and 3 red. The two sets of butterflies were identical except for their color. Two small butterflies in each color were 4.5 cm × 5.5 cm; one large butterfly in each color was 10 cm × 10 cm. We glued each set of butterflies onto wooden sticks, which were painted sky blue (20 cm). To display these stimuli, we constructed a box out of foam board. This box was rectangular with no top and with one side shorter than the other (see **Figure 1**, bottom left panel; front dimensions 30 cm × 33 cm, back dimensions 30 cm × 45 cm). The whole box was painted sky blue and white clouds were glued onto the inside to make it look like the butterflies were flying in the sky. We also used a commercially available sound device which could record and re-play a sound. This was placed at the back of the butterfly

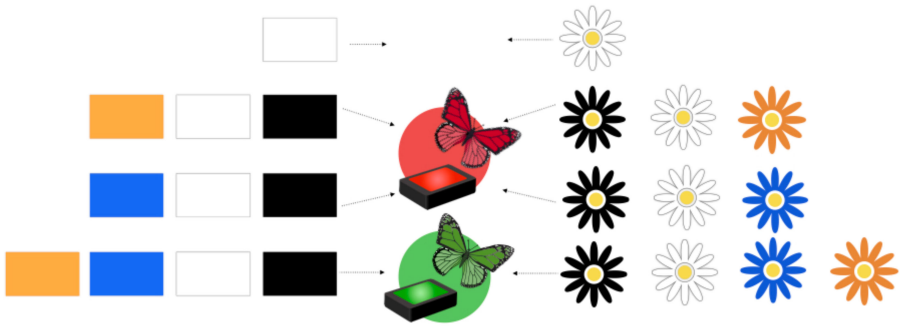


FIGURE 2 | Schematic of Order 1 for the blicket and butterfly tasks. Each row shows the combination of blocks/flowers used in one step of the demonstration phase of the task.

box where an experimenter could activate it out of view of the participant. Finally, as for the blicket version of the task, we made 7 laminated photographs of different combinations of the flowers (9.5 cm × 7.25 cm).

Both versions of the task also used a folded piece of cardboard as an occluder (approximately 60 cm × 100 cm) and 1 red and 2 green dots (1 cm in diameter).

Procedure

Figure 2 shows a schematic of the data that were presented to children in this study. Children saw four demonstrations of how the blicket detector system/flower system worked. These four demonstrations were shown in one of two orders. Order 1 (shown in **Figure 2**) first demonstrated the effect of the white block/flower; then the effect of the combination of the white, black, and orange blocks/flowers; then the effect of the white, black, and blue blocks/flowers; then the effect of all four blocks/flowers. Order 2 presented these combinations in the opposite order (all four, then white and black and blue, then white and black and orange, then white alone). Children were randomly assigned to an order. Here, we use Order 1 to illustrate the procedure.

Blicket condition

First, the experimenter placed the white block in the clear container and put the container on top of the machine. She narrated her actions: "Let's see what happens when I put just the white block on the machine". The machine made no response, which the experimenter noted aloud: "Nothing happened." She tried the white block again, again with no effect. She then brought out a photograph of the white block and put it on the table next to the machine, saying, "This is here to remind us what happened. When I put just the white block on the machine, nothing happened."

Second, she put the white, black, and orange blocks in the container and put the container on the machine. The machine turned red, and the experimenter noted this: "Look, it's turning red." (Note that our use of the container ensured that all of the blocks contacted the machine's surface at once, making it appear as though they were all jointly necessary for the effect.) The experimenter then tried the white, black, and orange blocks again, confirming that the machine turned red. She brought out the photograph of these three blocks and put a red dot next to it, saying, "Here's a red dot to remind us that the machine turned red when we put the white, black, and orange blocks on the machine." This same procedure was repeated for the combination of white, black, and blue blocks.

Finally, the experimenter put all four blocks in the container and set the container on top of the machine. This made the machine turn green and play music: "It's going green and playing music!" She repeated this demonstration a second time. Then she put down the photograph of all four blocks on the table. She put a green dot beside this photograph and said, "Here's a green dot to remind us that when we put the white, black, orange, and blue blocks on the machine, the machine turned green and made sound." The four photographs and their accompanying dots remained on the table for the rest of the procedure as a visual reminder to children of the data they had observed.

After showing all four of these combinations, the experimenter put up the cardboard occluder between the child and the machine in order to prevent the child from seeing what objects were placed on the machine. The experimenter told the child that she was placing two blocks on the machine and that the machine was turning green. The child could confirm that the machine turned green because it was playing the music that they had heard when the machine turned green before. The experimenter repeated this activation, again playing the music and saying, "It's turning green."

The experimenter then took away the occluder and reminded the child that the machine had just turned green when she put two blocks on it. She set out three new photographs, each of which showed a pair of blocks: blue and black, orange and black, and orange and blue. Then, the experimenter asked the main test question: "Which two blocks on the machine made it go green?" The correct answer is orange and blue: The white block has no effect, the pairs of blue/black and of orange/black were previously shown to turn the machine red, and the machine turned green only when the orange and blue blocks were on the machine together.

After the participant chose one of the options, the experimenter asked why they chose it. Note that, in this condition, this question was included mid-way through data

collection, so only 39 participants responded. Lastly, participants had the opportunity to try any combinations of blocks they wanted on the machine and to observe the effects.

Butterfly condition

This condition paralleled the procedure in the Blicket condition, using flowers instead of blocks and butterflies instead of lights. That is, instead of putting blocks in a container and placing the container on a machine to make it turn red or green, the experimenter "planted" different combinations of flowers in the foam block ("field") and then placed the block next to the light blue box ("sky"). Instead of showing red or green lights, the experimenter made either the red or the green butterflies appear by lifting them on their stick to poke out above the front wall of the box apparatus (see bottom center panel of **Figure 1**). When the green butterflies appeared, she also pushed the button on the sound machine to make a song play. As in the Blicket condition, participants were provided with photographs of the four combinations of flowers, paired with green or red dots, as a visual reminder of what they had observed.

Participants were shown the same combinations of data and effects, in either Order 1 or Order 2 (see **Figure 1**). After observing the four demonstrations, the experimenter put up the occluder and said that she was planting two flowers in the field, which brought green butterflies. She activated the sound machine to make the sound that has previously been associated with the appearance of the green butterflies, so that participants could verify her claim, as in the Blicket condition. Then, they were asked the same test question as in the Blicket condition: "Which two flowers made the green butterflies come?" After responding to this question, all participants in this condition were asked to justify their response. Finally, as an exploratory investigation of children's own actions on causal systems, they were also given the opportunity to plant whatever combination of flowers they wanted and observe the effects.

There were 116 children in the blicket condition and 126 children in the butterfly condition; there are fewer children in the blicket condition because errors with the detector led to more exclusions in this condition. Condition assignment was only partially random because we began by running most of the blicket condition first, which explains the discrepancy in the number of participants who were asked the justification question by condition. Participants in both conditions also engaged in an open-ended interview about the word "science" as part of a different investigation. Their responses to this question had no bearing on their performance in either the Blicket or Butterfly conditions and will not be discussed further.

Coding

In both conditions, we coded answers of blue/orange as correct and answers of black/orange or black/blue as incorrect. Children's justifications were coded as to whether they were relevant to the task or irrelevant. Relevant justifications appealed to the data that the child had observed in the demonstration phase or to the fact that a particular combination of colors would be efficacious (e.g., "because gray and orange were in this when it made green"). Justifications that reflected the child's beliefs but that did not provide any substantive information (e.g., "I just think it be the right ones,") were coded as irrelevant. Justifications of "I don't know" or cases in which children did not provide a justification were also coded as irrelevant. Justifications were coded by two naïve research assistants, blind to children's age and gender. Agreement was 90% (Kappa = 0.79). Disagreements were resolved by discussion with the third author.

Results

Table 1 shows performance on both the blicket and butterfly tasks across the age groups. Preliminary analysis showed that order of trials children received did not affect performance, $\chi^2(1, N = 242) = 0.03$, $p = 0.85$, so we did not analyze this factor further. There was also not a significant difference between girls' and boys' responses, $\chi^2(1, N = 242) = 2.67$, $p = 0.11$, so we did not consider this factor further. To analyze responses to the main test question, we constructed a general linear model with a binomial logistic distribution, analyzing the role of age (in months) and condition (blicket, butterfly) using a main effect only model (which proved more explanatory than a factorial model as evidenced by a lower BIC value[1]). The overall model was significant, $\chi^2(2) = 11.33$, $p = 0.003$. A main effect of age was found, Wald $\chi^2(1) = 10.56$, $p = 0.001$. No effect of condition was found, Wald $\chi^2(1) = 0.23$, $p = 0.63$.

We broke the two conditions into three age groups shown in **Table 1**. These groups roughly corresponded to 4- and 5-year-olds, 6- and 7-year-olds, and 8–10-year-olds. In both conditions, the two younger groups responded no different from chance (33%, because we presented 3 answer choices), all Binomial tests, $p > 0.19$. The older group responded differently from chance, Binomial tests, $p = 0.007$ in the blicket condition and $p < 0.001$ in the butterfly condition.

Justifications

As noted above, only 39 of the children in the blicket condition were asked to justify their responses (compared to 125 of

[1]We use this strategy throughout the analysis of all three experiments. We considered models with main effects only and models that included both main effects and interactions. We only report the latter if it was a better fitting model. We indicate where this is the case; in all other cases throughout these experiments, the main effects only model was the better-fitting model.

TABLE 1 | Proportion of correct responses in each condition by age groups in experiment 1.

	Youngest third	Middle third	Oldest third
Blicket Condition	38.46	41.03	55.29
	(49.29)	(49.83)	(50.39)
	$n = 39$	$n = 39$	$n = 38$
	M_{age} = 58.84 months	M_{age} = 84.65 months	M_{age} = 108.17 months
Butterfly condition	26.19	38.10	59.52
	(44.50)	(49.15)	(49.68)
	$n = 42$	$n = 42$	$n = 42$
	M_{age} = 58.76 months	M_{age} = 80.44 months	M_{age} = 108.44 months

the children in the butterfly condition), because this question was added partway through data collection. Fifteen children in the blicket condition (38.5%) and 38 children in the butterfly condition (30.4%) provided relevant justifications. This was not a significantly different ratio, $\chi^2(1, N = 164) = 0.35$, $p = 0.35$. Additionally, 9 children in the blicket condition (23.1%) and 25 in the butterfly condition (20.0%) said "I don't know," said that they were guessing, or made no response to this question; this response tendency also did not significantly differ by condition $\chi^2(1, N = 164) = 0.17$, $p = 0.68$.

To examine any potential links between children's justifications and their performance, we conducted a binary logistic regression on whether children generated a relevant justification based on their age, condition, and whether they responded correctly on the test question. The overall model was significant, $\chi^2(3) = 27.08$, $p < 0.001$. Age predicted a unique proportion of variance, $\beta = 0.04$, SE = 0.01, Wald $\chi^2 = 16.85$, $p < 0.001$. Condition and performance on the test question were not significant in this model.

Open-Ended Experimentation

After children justified their choice, they were allowed to try any combination of blocks/flower on the machine/in the field. Because different children tried different numbers of combinations, we only examined children's first attempts. Our primary goal with this question was to explore whether children chose the correct combination of orange and blue, regardless of whether they had chosen this pair at test. Although this task was not designed to directly probe children's predictive abilities, we speculated that this response tendency could indicate some implicit understanding of how the machine worked, even in the absence of a correct explicit response (see e.g., Cook et al., 2011).

Ninety-five children in the blicket condition and 119 children in the butterfly condition tried at least one combination of stimuli. Of these children, 38% in the blicket condition while only 24% in the butterfly condition tried the combination orange and blue (i.e., the correct response to the test question), a significant difference, $\chi^2(1, N = 214) = 5.20$, $p = 0.02$, Phi = −0.16.

We performed a binary logistic regression examining whether children tried the correct pair of stimuli, looking at age, condition, and whether children chose the correct response on the test trial. The overall model was significant, $\chi^2(3) = 12.72$, $p = 0.005$. Condition explained unique variance, with children in the Blicket condition being more likely than children in the Butterfly condition to try the orange/blue combination, $\beta = -0.66$, SE = 0.31, Wald $\chi^2 = 4.63$, $p = 0.03$. Correct responding on the test trial also explained unique variance; children who responded correctly on the test trial were significantly more likely to choose the orange/blue pair to test in their open-ended exploration, $\beta = 0.86$, SE = 0.32, Wald $\chi^2 = 7.38$, $p = 0.007$. Age did not explain unique variance, $\beta = 0.006$, SE = 0.007, Wald $\chi^2 = 0.56$, $p = 0.45$. We then ran a separate model with only condition, correct responding and the interaction between them. This model was a better fit (as indicated by a lower BIC value, 261.38 vs. 264.88), and indicated that the interaction between condition and correct responding was also significant, $\beta = 1.26$, SE = 0.63, Wald $\chi^2 = 3.99$, $p = 0.046$. Specifically, in the Blicket condition, children were more likely to try the orange/blue pair if they had chosen it at test (56% vs. 23% of the time). This difference was not as great in the Butterfly condition (25% vs. 22%).

Discussion

Four- to 10-year-olds were given one of two kinds of diagnostic reasoning tasks. The first was a replication of the procedure used by Sobel et al. (2017), in which children had to make inferences about an interactive causal structure. The second was an inferential problem with the same causal structure, but different content – about biological instead of physical mechanisms. We generally replicated the earlier results using the blicket detector, and showed that children generated responses to the butterfly version of the same problem with a similar developmental trajectory. In many ways, this investigation parallels findings by Schulz and Gopnik (2004), who showed that preschoolers use the same domain-general formal principles of causal inference when reasoning about physical and biological content.

Unlike children's performance with the main test question, the combinations of blocks that children chose to try in response to our open-ended prompt differed between the two conditions: Children in the Blicket condition were more likely to choose the correct (orange/blue) combination. Children in this condition who chose to try the correct combination were also more likely to have chosen this combination at test than in the Butterfly condition, potentially implying that they were verifying their choice. Because these choices reflect children's predictive (rather than diagnostic) abilities, which were not the focus of this investigation, we hesitate to draw any strong conclusions about this result.

Finally, due to a change in procedure, only a subset of children in the Blicket condition in Experiment 1 was asked to justify their responses. We thought it critical to replicate this procedure to ensure the robustness of our findings about children's justifications. Experiment 2 does so, while also presenting the two conditions as a within-subject manipulation. This allowed us to try to replicate the findings of Experiment 1 while controlling for individual-level variance and examining whether children notice the similarities between the two tasks.

EXPERIMENT 2

Method

Participants

The final sample included 103 children between the ages of 4–10 (55 girls, 48 boys, $M_{age} = 82.93$ months, $SD_{age} = 22.33$ months) from preschools, elementary schools, and after-school programs. As there is not an agreed-upon protocol for *a priori* power analysis for GEE, we aimed to test as many children as we could at our participating schools and to roughly match the sample size of Experiment 1. Twelve additional participants were tested, but not included: One did not respond to any questions, and for the other eleven, there were errors with the activation of the machine. In terms of race, 53 families identified as White, 34 identified as African American, 7 identified as multiracial, 1 identified as

American Indian or Alaskan Native, 2 identified as Asian, and 6 did not provide this information. In terms of ethnicity, 9 families identified as Hispanic or Latino, 31 identified as not Hispanic or Latino and 63 did not provide this information.

Materials and Procedure

The same materials used in Experiment 1 were used in Experiment 2. The only difference in procedure was that Experiment 2 ran the two tasks one after the other (order counterbalanced). Additionally, in Experiment 2, we asked all participants to justify their answer choices. Finally, following children's completion of their second task, we added a question to probe for whether they noticed anything similar about the two tasks: "We just played the machine [butterfly] game. Do you remember before when we played the butterfly [machine] game? Was anything the same about the machine game and the butterfly game?" We did this to investigate whether any children would code the similarity between the tasks as being deeper than just surface-level, and if their noticing of structural similarities between the tasks might relate to better performance.

Coding

We used the same coding for the main test questions, children's justifications, and children's first open-choice combination as in Experiment 1. Coding for the justifications was again done by two research assistants, blind to children's age and gender, but not blind to whether they were coding a blicket or butterfly response (coding for the blicket trials was done at a separate time than coding for the butterfly trials). Agreement for the butterfly trials was 91% (Kappa = 0.81). Agreement for the blicket trials was 95% (Kappa = 0.89). Disagreements were resolved through discussion with the second author.

For the similarity question, we scored children's responses as 1 if they said they noticed that something was the same about the two tasks and 0 otherwise. For children who recognized the similarity, we coded their reasons as either *causal* or *perceptual*. Causal responses mentioned the similarity of causal structure of the two games, for example, "both of them turned green when you put all of them." Perceptual responses mentioned the similarity of any perceptual aspect of the two games, for example, "both were red and green." Two coders independently coded 10% of the sample and agreement was 100%. A single coder then coded the remainder of the sample.

Results

Preliminary analyses revealed no effects of order (i.e., receiving the blicket or butterfly version first) on responses to either test question, both $\chi^2(1, N = 103)$-values < 0.47, both p-values > 0.61. Similarly, there was no difference between girls and boys in their response to either test question, $\chi^2(1, N = 103)$-values < 0.70, both p-values > 0.41. As a result, these variables were not considered further. We constructed a General Estimating Equation (GEE) with an independent working correlation matrix and a binomial distribution (following Zeger and Liang, 1986; Zeger et al., 1988). This accounted for the within-subject nature of our procedure. Correct performance was the dependent variable, and age and condition were the independent variables. This model revealed a significant main effect of age, Wald $\chi^2 = 19.86$, $p < 0.001$, but not a significant effect of condition, Wald $\chi^2 = 0.33$, $p = 0.56$.

TABLE 2 | Proportion of correct responses in each condition by age groups in experiment 2.

	Youngest third (n = 34)	Middle third (n = 35)	Oldest third (n = 34)
Blicket condition	23.53 (43.06)	42.86 (50.21)	70.59 (46.25)
Butterfly condition	29.41 (46.25)	37.14 (49.02)	58.82 (49.96)

Overall, the sample's performance on the blicket measure (46% accurate) was better than what would be expected by chance (33%), Binomial test, $p = 0.005$, as was performance on the butterfly measure (42%), Binomial test, $p = 0.04$. Given the results of Experiment 1, we separated our sample into three roughly equal groups based on age. These results are shown in **Table 2**. The youngest group (roughly 4–5-year-olds, M_{age} = 57.03 months, Range 46.20–68.63 months) responded no different from chance on either measure, Binomial tests, both p-values > 0.15. The middle group (roughly 6–7-year-olds, M_{age} = 83.89 months, Range 69.57–96.20 months) also responded no different from chance on either measure, Binomial tests, both p-values > 0.15. The oldest group (roughly 8–10-year-olds, M_{age} = 107.84 months, Range 96.93 132.87 months) responded above chance on both measures, Binomial tests, $p = 0.002$ for the butterfly measure and $p < 0.001$ for the blicket measure.

Justifications

Children generated relevant justifications on 37% of the butterfly trials and 31% of the blicket trials. A new General Estimating Equation was built to analyze appropriate responses on each trial, with age, condition, and whether children responded correctly on the test question as dependent variables. This model revealed a main effect of age, Wald $\chi^2 = 5.25$, $p = 0.02$. Condition was not significant, Wald $\chi^2 = 1.55$, $p = 0.21$. The effect of correct response on the test trials did not reach significance, but children who had responded correctly were marginally more likely to generate relevant justifications, Wald $\chi^2 = 3.00$, $p = 0.08$.

Open-Ended Experimentation

We again examined how many children selected the orange/blue (correct) combination as the first combination they wanted to try. To parallel the analysis from Experiment 1, we ran a GEE on whether children tried the correct combination in each trial, with age, condition, and correct responding as dependent variables. This resulted in a main effect of condition, $\beta = -0.98$, SE = 0.35, Wald $\chi^2 = 8.01$, $p = 0.005$ and a main effect of correct responding, $\beta = 1.20$, SE = 0.39, Wald $\chi^2 = 9.32$, $p = 0.002$, but not a main effect of age[2]. Of importance is that the main effect of condition was the opposite of Experiment 1. Here, for the blicket condition, 13% of children's attempts were orange/blue, compared to 25% for the butterfly condition. This was a significant difference in ratios, McNemar $\chi^2(1, N = 103) = 6.50$, $p = 0.01$.

[2]Because children who responded incorrectly on the test question never tried the correct combination on the machine, it was not possible to test the interaction, which was significant in Study 1.

In terms of the effect of correct responding, in the butterfly condition, children who responded correctly on the test question were more likely to choose to try the correct combination in the butterfly condition (33%) than children who did not respond correctly (20%). Similarly, in the blicket condition, children who responded correctly on the test question were more likely to choose to try the correct combination (29%) than children who did not (0%). Only in the blicket condition did this difference reach statistical significance, Fisher Exact Test, $p = 0.18$ and $p < 0.001$.

Similarity Question

Three children did not respond to the question about similarity. Of those who did, 82% of the participants said that they noticed some similarity between the two tasks, significantly more than chance, Binomial Test, $p < 0.001$. This response did not correlate with children's age, $r_s(98) = 0.10$, $p = 0.32$. When asked to justify why children believed the two trials were similar, 70% mentioned perceptual similarity whereas 30% generated a causal justification. Generating a justification of a particular type did not significantly correlate with age, nor did it correlate with performance on either the blicket or butterfly trial, all $r_s(80)$-values $< |0.14|$, all p-values > 0.23.

Comparisons to Experiment 1

Finally, we tested whether children responded differently on either the butterfly or the blicket conditions between Experiment 1 and 2. We found no difference in performance on either the blicket conditions across the studies (46% in each study), $\chi^2(1) < 0.001$, $p = 0.99$, or on the butterfly conditions (42% in each study), $\chi^2(1) = 0.001$, $p = 0.98$. This suggests that the within-subject nature of this experiment did not affect performance compared to a between-subject version of the same measure.

Discussion

Experiment 2 found similar results to Experiment 1, whereby performance on both tasks improved with age, though only children in the oldest group performed at above-chance levels. We also found no differences between performance on these two tasks for any age group, either in terms of their responses to the test questions or in terms of their justifications for these responses. As in Experiment 1, 4- to 7-year-olds mostly responded at chance levels of performance, while 8- to 10-year-olds were generally above chance at making the appropriate diagnostic inference.

When examining which combinations of blocks children chose to try in the two causal systems, we did find a difference between the conditions, but in the opposite direction from Experiment 1. Here, children were more likely to try the correct answer in response to our open-ended prompt (orange/blue) for the Butterfly task than for the Blicket task. Additionally, for both tasks (though only statistically significantly for the Blicket task), children were more likely to try the orange and blue blocks if they had previously chosen this pair at test. Future work should look more directly at this relation between children's diagnostic and predictive reasoning, as the current studies were not designed to specifically capture this aspect of children's thinking, and past work suggests that these two reasoning process are not symmetrical (see e.g., Kuhn et al., 2009; Fernbach et al., 2010).

Finally, for the question about similarity between the two procedures, most of the children (correctly) said that the two tasks were similar to each other, although only a minority of children articulated that this similarity was due to the underlying causal structure rather than to the perceptual features of the two tasks. Saying that the two tasks were similar to each other did not affect performance on either measure, nor did children's reported reason for why the tasks were similar (for example, talking about the causal structure or the more superficial perceptual features).

The two experiments so far suggest that children show similar diagnostic reasoning abilities across different domains of content. However, the blicket and butterfly measures were equivalent in a variety of ways, beyond just the causal structure. Both tasks involved diagnosing a set of causal relations, and among those causal relations, children had to track different levels of a single feature (color) to discern among potential causes. While the causal system was novel, in neither condition did children hear any information that was potentially unfamiliar or that required additional definitions or background knowledge – that is, the potential causes were all easily identifiable and comprehensible. All of these factors potentially facilitated children's diagnostic inference.

EXPERIMENT 3

In Experiment 3, we introduced a novel measure that used the same interactive structure used in Experiments 1 and 2, but that couched the problem in a more realistic scientific framework, increasing the information processing demands of the task. Instead of asking children to use colors make inferences about a set of causal relations, we asked children to use a set of biological features to infer category membership (whether an exemplar was a particular kind of dinosaur). The four features were presented as an interactive structure: Having both feature A and B meant that the example was one kind of dinosaur; having only A or only B meant that the example was a second kind; and having neither meant that the example was a third. However, each feature was unique (unlike Experiments 1 and 2, which used different levels of the color feature) and not directly observable (as was the case with the individual blocks or flowers). Thus, although Experiment 3 in some ways presents the same structure as Experiments 1 and 2, it differs in many important ways, which serve to make it a better test of the kinds of diagnostic inferences that children are asked to make in classrooms and in real life.

The fundamental question behind our investigation so far has been to examine whether children engage in diagnostic inference differently across domains. The first two experiments support the hypothesis that there is little difference between the inferences that children can make in the physical and biological domains. Biological relations, however, can be much more complex than what we manipulated in Experiments 1 and 2. Experiment 3 thus presents a more stringent test of our research question.

Method

Participants

The final sample included 110 children between the ages of 4–10 (52 girls, 58 boys, M_{age} = 80.08 months, SD_{age} = 22.62 months), again aiming to roughly match the sample size of Experiment 1. These children were primarily recruited from and tested in local museums (n = 107), and a minority were tested at our lab or in a preschool (n = 3). We additionally tested 2 participants who were not included in the final sample. One child was a 3-year-old; the other refused to respond to questions.

The demographics of this sample were as follows: 68 families identified as White, 9 identified as African American, 3 identified as multiracial, 7 identified as Asian, and 23 did not provide this information. In addition, 12 families identified as Hispanic or Latino, 13 identified as not Hispanic or Latino and 85 did not provide this information.

Materials

We used a whiteboard (22.75 inches wide x 17 inches high) to present a grid that showed which combinations of traits were characteristic of which kinds of dinosaurs (see **Figure 3** and the Procedure section for more details). We created a set of 8 laminated pictures, one for each of four traits (e.g., having a beak-shaped mouth) and one for each of four dinosaurs (e.g., Einiosaurus). These could be stuck to the whiteboard using Velcro dots during the procedure to provide a visual aid for our explanations. We also used a green and a red whiteboard marker to indicate the presence (green check) or absence (red X) of each trait.

Procedure

To create the dinosaur version of the diagnostic reasoning task, we selected three dinosaurs from the family of ceratopsians (horned-headed dinosaurs): Einiosaurus, Triceratops, and Zuniceratops. We also selected four traits that these dinosaurs could have: mouth shape (beak-shaped or not), size (larger or smaller than a human when fully grown), head crest (large or small), and horn direction (backwards-facing or forwards-facing). As in the other two versions of this task, children saw four demonstrations of how different traits combined to characterize the three types of dinosaurs. These demonstrations were presented in one of two orders. Order 1 first presented Einiosaurus, then adult Triceratops, then baby Triceratops, then Zuniceratops; Order 2 presented the dinosaurs in the reverse order. We here use Order 1 to illustrate the procedure.

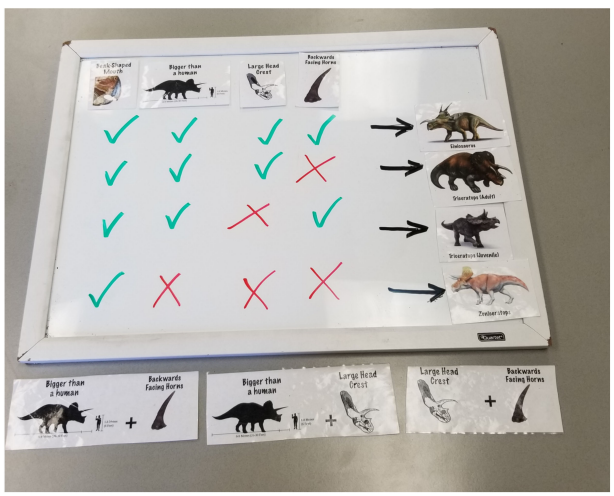


FIGURE 3 | Completed grid for the dinosaur task. The experimenter would first explain and place the traits as column headers, then add checks and Xs with whiteboard markers to characterize the dinosaurs. The three answer choices appear at the bottom.

Children were told that they would play a dinosaur game to learn about the different traits that dinosaurs can have. They were then oriented to the grid on the whiteboard and to the four traits that served as column headers (see **Figure 3**). An experimenter first told children that some dinosaurs belong to the ceratopsian family, which means that they have beak-shaped mouths. The experimenter showed the child the picture card that illustrated a beak-shaped mouth, then stuck it to the first column on the whiteboard. The same was done for the next three traits: being larger than a human when fully grown, having a large head crest, and having backwards-facing horns. As the experimenter described each trait, she showed the child a picture of that trait and then stuck it to the whiteboard to fill out the column headers. These traits were always presented in the same order.

Then, children were told that they would see some dinosaurs, and that some of these dinosaurs have these traits while others do not. Order 1 first introduced Einiosaurus. The experimenter brought out the picture card for this dinosaur and explained that Einiosaurus has all four traits: It has a beak-shaped mouth, is larger than a human when fully grown, has a large head crest, and has backwards-facing horns. This dinosaur was thus the parallel of placing all four blocks on the machine or planting all four flowers in the field. As the experimenter named each of the four traits, she used the green whiteboard marker to put a green check under each column of the first row of the grid on the whiteboard. At the end of this introduction, the experimenter repeated the combination of traits: "So if a dinosaur has a beak-like mouth and is bigger than a human when it's fully grown and has a big head crest and has backwards facing horns, that means it is an Einiosaurus." She then stuck the Einiosaurus picture card to the right of the first row.

The same procedure was repeated for the other three dinosaurs, using green checks when a dinosaur had a trait and a red Xs when a dinosaur lacked a trait (see **Figure 3**). Adult Triceratops has a beak-shaped mouth, is larger than a human when fully grown, and has a large head crest, but it does not have backwards-facing horns; its horns face forward. Baby Triceratops has a beak-shaped mouth, is larger than a human when fully grown, and has backwards-facing horns, but it has a small head crest. Because Triceratops changes its head crest size and horn direction as it matures (Horner and Goodwin, 2006), it served as the equivalent of the red activation on the blicket machine or the red flowers in the butterfly field: Two different combinations of three traits led to the same type of dinosaur, just as two different combinations of three blocks/flowers led to a red light/red butterflies.

The fourth dinosaur, Zuniceratops, has only one of the four traits: a beak-shaped mouth. But it is smaller than a human when fully grown, lacks a head crest, and has forward-facing horns.

At the end of this demonstration phase, children could see the full set of trait combinations on the grid on the whiteboard (**Figure 3**). To frame the test question, children were told that paleontologists have found a new set of fossils that they are sure is an Einiosaurus. But the paleontologists used only two traits to figure that out. Children were asked to choose which pair of traits the paleontologists used to know for sure that the fossils they found belonged to an Einiosaurus: (a) being bigger than a human and having backwards-facing horns, (b) being bigger than a human and having a large head crest, or (c) having backwards-facing horns and having a large head crest. In parallel to the other two versions of this task, the correct answer is c, since only the Einiosaurus has both of these traits.

These answer choices were accompanied by illustrations that combined the relevant pictures from the column headers on the whiteboard (see bottom of **Figure 3**). The answer choices were presented in a random order for each child. The corresponding picture cards were placed on the table in front of the child as the experimenter described each one. As in the previous versions of this task, after choosing a response option, children were asked to justify their choice.

Coding

Answers of "backwards-facing horns and having a large head crest" were coded as correct; the other two answers were coded as incorrect. We coded children's justifications as *relevant* when they mentioned the different traits that the dinosaurs could have or when they made comparisons between the dinosaurs (e.g., "well, if you think it's this [pointing to Einiosaurus] bigger than a human, it's pretty easy, and a large headcrest is only on two of them, which I've seen so that makes it, so that means narrowing the options and so from there we can just look at the details and probably try to figure it out."). We coded children's justifications as *irrelevant* when they did not provide any substantive information (e.g., "Because it is bigger than a human and because dinosaurs are so big") or when they said "I don't know." Justifications were coded by two research assistants who were blind to children's age and gender. These two coders considered 89% of the dataset. Agreement was 89% (Kappa = 0.77). Disagreements were resolved through discussion with the third author and then the first coder coded the rest of the data.

Results

Preliminary analyses showed that the order in which information was presented did not affect responses, $\chi^2(1, N = 110) = 0.04$, $p = 0.85$. Preliminary analyses also showed that girls and boys did not differ in their responses, $\chi^2(1, N = 110) = 0.01$, $p = 0.94$. As a result, we do not consider these factors further.

Overall, children responded correctly 27% of the time, not significantly greater than chance, Binomial test, $p = 0.20$. We tested for age effects using a general linear model with a binomial logistic distribution, as in Experiment 1. The overall model was not significant, $\chi^2(1) = 3.17$, $p = 0.08$, nor was the individual effect of age, B = −0.02, SE = 0.01, 95% CI [−0.04, 0.00], Wald $\chi^2(1) = 3.11$, $p = 0.08$. As a way of initially contrasting performance in this condition to that of the other conditions, we broke the sample into thirds based on age, as we did in the previous two studies. No age group's performance was above chance responding (youngest: 25%; middle: 21%; oldest: 36%; all Binomial tests, p-values > 0.14).

We examined children's justifications by conducting a general linear model with a binary logistic distribution, predicting whether children generated a relevant justification based on their age and whether they responded correctly on the test question. The overall model was significant, $\chi^2(2) = 10.35$, $p = 0.006$. Age was a significant predictor, Wald $\chi^2(1) = 9.14$, $p = 0.003$. Correct judgments on the test question was not a significant predictor, Wald $\chi^2(1) < 0.01$, $p = 0.99$.

To contrast performance with the blicket and butterfly conditions in Experiment 1, we again used a general linear model with a binomial logistic distribution, examining the role of age (in months) and task version (blicket, butterfly, dinosaur). The overall model was significant, $\chi^2(3) = 22.54$, $p < 0.001$. This analysis revealed a main effect of condition, Wald $\chi^2(2) = 7.12$, $p = 0.03$ and a main effect of age, Wald $\chi^2(1) = 13.60$, $p < 0.001$. Overall, performance on the dinosaur condition (27% correct) was worse than performance on the blicket condition (45% correct), B = −0.74, SE = 0.29, 95% CI [−1.30, −0.17], Wald $\chi^2(1) = 6.46$, $p = 0.01$, and performance on the butterfly condition (41%), B = −0.61, SE = 0.29, 95% CI [−1.17, −0.05], Wald $\chi^2(1) = 4.52$, $p = 0.03$.

Discussion

On this dinosaur task, we found no improvement in performance with age, and even the oldest age group tested responded at chance levels. Although the underlying reasoning structure of the dinosaur task was identical to that of the blicket and butterfly tasks presented in Experiments 1 and 2, children had more difficulty solving this version of the problem. As prior work has shown that children understand the role of biological features in categorization (Keil, 1989; Carey, 1995; Slaughter and Lyons, 2003), we believe that this difference in performance is due to other features of our task, most importantly the interaction between the task's surface features and its underlying structure: The use of dinosaurs made it appear as though domain-specific knowledge was necessary to solve the task, although this was not the case. We discuss this issue further in the General Discussion, recalling that our goal with this experiment was to begin to explore how more realistic science content can affect children's reasoning, and not to investigate the impact of individual task features on children's performance.

GENERAL DISCUSSION

The current study examined whether children possessed similar diagnostic reasoning abilities across different domains of knowledge. In Experiment 1, we replicated previous findings that suggested children could engage in diagnostic reasoning about an interactive causal structure by age 7 (Sobel et al., 2017). We

also extended this finding to another domain of knowledge: biology. We demonstrated that children's diagnostic reasoning is not limited to the blicket detector paradigm; children performed similarly on both versions of the task. Using a within-subject design, instead of a between-subject design, Experiment 2 replicated the findings of Experiment 1, as well as corrected for an error in data collection regarding children's justifications of their inferences. These results imply that the mere presence of domain-specific features does not impact children's abilities to reason diagnostically. Further, these results confirm that this kind of diagnostic reasoning emerges around age 7, possibly because younger children lack the capacity to coordinate the high level of uncertainty about the efficacy of individual variables presented by our task (e.g., Erb and Sobel, 2014) or lack the metacognitive capacities to reflect on this uncertainty (Kuhn and Dean, 2004).

In contrast to the first two experiments, children performed much more poorly when asked to make diagnostic inferences about category membership of dinosaurs based on their features (Experiment 3). Here, although the underlying structure of the problem was the same as in Experiments 1 and 2, children did not show improvement with age, and even the oldest children we tested did not respond correctly above chance levels.

There were, of course, many important differences between the dinosaur version of the task presented in Experiment 3 and the tasks presented in Experiments 1 and 2, any of which (or their combination) could explain this difference in performance. Children in Experiment 3 had to track different features (e.g., size and shape) as relevant for category membership, as opposed to different levels of a single feature (color). In addition, there was a difference in degree of complexity of the outcomes: The dinosaur condition featured children learning the names of different (potentially novel) dinosaurs and their features, whereas the other conditions did not require the child to learn novel vocabulary or other scientific content. Children may also have differed in their level of domain-specific knowledge about biology in general or dinosaurs in particular, which may have affected the degree to which they believed that the task required such knowledge to solve. Future work should attempt to examine which of these features have the most impact on children's reasoning abilities. But, crucially, it might not be possible for such investigations to disentangle whether children experience more difficulty in more realistic problems because of information processing demands or because of the realism itself; increasing the amount of domain-specific knowledge in a domain of reasoning will necessarily involve adding information to a given causal model.

Equally crucially, we would argue that the fact that the dinosaur version of the task was not precisely matched to the blicket and butterfly versions of the task is somewhat beside the point. Our goal in Experiment 3 was to instantiate a causal structure for which the developmental trajectory was well-understood (from work with blicket detectors) in a context that superficially resembled reasoning tasks that children see in classrooms or in prior work on diagnostic inferences (Kuhn and Dean, 2005; Kuhn et al., 2009). In doing so, although we modified more than just one aspect of the task, we were able to take an initial step toward reconciling conflicting conclusions from cognitive developmental psychology, claiming that even babies can reason diagnostically (e.g., Gopnik et al., 2004; Sobel and Kirkham, 2006), with work from education science, claiming that children struggle with diagnostic thinking until late elementary school and beyond (e.g., Klahr et al., 1993; Chen and Klahr, 1999).

In particular, the differences demonstrated here between the blicket and butterfly versions of the reasoning task and the dinosaur version have important implications for how the development of diagnostic reasoning is understood. Children's performance on the blicket and butterfly versions of the task demonstrated that at least the older children in our sample (8- to 10-year-olds) have the reasoning skills necessary to diagnose the interactive causal system that we presented [as also shown by Koerber et al. (2015) and Sobel et al. (2017)]. On this basis, some have suggested that these children (and even younger children) are mature scientific thinkers, since diagnostic reasoning is an important component of scientific thinking (e.g., Cook et al., 2011; Gopnik and Wellman, 2012).

But there are many important methodological differences between the tasks used in those studies (and in our Experiments 1 and 2) and other investigations of scientific reasoning (and in our Experiment 3), such as understanding experimental design, engaging in hypothesis formation, and reflecting on the quality of data and what inferences those data support. Many of these differences have been discussed and investigated elsewhere (e.g., Sodian et al., 1991; Ruffman et al., 1993; Kuhn and Pearsall, 2000; Kuhn, 2007; Sobel et al., 2017) and are beyond the scope of the present investigation. Indeed, these additional reasoning capacities are likely only available to older children, and the continuing developmental trajectory of children's scientific reasoning abilities may involve increasing coordination among these capacities and the kind of diagnostic reasoning abilities investigated in the current work. Rather, we wish to emphasize that prior work on scientific reasoning has tended to focus on making diagnostic inferences about domain-specific contexts, like those found in science classrooms (see e.g., NGSS Lead States, 2013). Because our Experiment 3 resembles such systems more closely, our results suggest that the information processing demands of such contexts (which come from many sources) may prevent children from demonstrating their existing capacities to reason about such systems. Children may have broad diagnostic reasoning capacities in the early elementary school years, but applying those capacities to the science classroom seems to only come later in development.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Villanova University Institutional Review Board.

Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

DW and DS conceptualized the study and secured funding. EC executed the design, collected the data for Studies 1 and 2, and wrote the first draft of the manuscript. DW executed the design and collected the data for Study 3. DS performed the statistical analyses. All authors then contributed to the editing process.

FUNDING

This research was supported by the National Science Foundation (NSF BCS-1929935 to DW and NSF BCS-1661068 to DS).

REFERENCES

Beck, S. R., Robinson, E. J., Carroll, D. J., and Apperly, I. A. (2006). Children's thinking about counterfactuals and hypotheticals as possibilities. *Child Dev.* 77, 413–423. doi: 10.1111/j.1467-8624.2006.00879.x

Bonawitz, E. B., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E. S., and Schulz, L. E. (2011). The double-edged sword of pedagogy: instruction limits spontaneous exploration and discovery. *Cognition* 120, 322–330. doi: 10.1016/j.cognition.2010.10.001

Carey, S. (1995). "On the origin of causal understanding," in *Causal Cognition: A Multidisciplinary Debate*, eds D. Sperber, D. Premack, and A. J. Premack (Oxford: Clarendon Press), 268–308.

Chen, Z., and Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Dev.* 70, 1098–1120. doi: 10.1111/1467-8624.00081

Cook, C., Goodman, N. D., and Schulz, L. E. (2011). Where science starts: spontaneous experiments in preschoolers' exploratory play. *Cognition* 120, 341–349. doi: 10.1016/j.cognition.2011.03.003

Erb, C. D., and Sobel, D. M. (2014). The development of diagnostic reasoning about uncertain events between ages 4–7. *PLoS One* 9:e92285. doi: 10.1371/journal.pone.0092285

Fernbach, P. M., Darlow, A., and Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychol. Sci.* 21, 329–336. doi: 10.1177/0956797610361430

Fernbach, P. M., Darlow, A., and Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *J. Exp. Psychol. Gen.* 140, 168–185. doi: 10.1037/a0022100

Fernbach, P. M., Macris, D. M., and Sobel, D. M. (2012). Which one made it go? The emergence of diagnostic reasoning in preschoolers. *Cogn. Dev.* 27, 39–53. doi: 10.1016/j.cogdev.2011.10.002

Goodman, N. D., Ullman, T. D., and Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychol. Rev.* 118, 110–119. doi: 10.1037/a0021336

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychol. Rev.* 111, 3–32. doi: 10.1037/0033-295X.111.1.3

Gopnik, A., and Sobel, D. M. (2000). Detecting blickets: how young children use information about novel causal powers in categorization and induction. *Child Dev.* 71, 1205–1222. doi: 10.1111/1467-8624.00224

Gopnik, A., Sobel, D. M., Schulz, L. E., and Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Dev. Psychol.* 37, 620–629. doi: 10.1037/0012-1649.37.5.620

Gopnik, A., and Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychol. Bull.* 138, 1085–1108. doi: 10.1037/a0028044

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., and Gopnik, A. (2011). Bayes and blickets: effects of knowledge on causal induction in children and adults. *Cogn. Sci.* 35, 1407–1455. doi: 10.1111/j.1551-6709.2011.01203.x

Horner, J. R., and Goodwin, M. B. (2006). Major cranial changes during Triceratops ontogeny. *Proc. R. Soc. B Biol. Sci.* 273, 2757–2761. doi: 10.1098/rspb.2006.3643

Keil, F. C. (1989). *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Klahr, D. (2002). *Exploring Science: The Cognition and Development of Discovery Processes*. Cambridge, MA: MIT Press.

Klahr, D., Fay, A. L., and Dunbar, K. N. (1993). Heuristics for scientific experimentation: a developmental study. *Cogn. Psychol.* 25, 111–146. doi: 10.1006/cogp.1993.1003

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., and Sodian, B. (2015). The development of scientific thinking in elementary school: a comprehensive inventory. *Child Dev.* 86, 327–336. doi: 10.1111/cdev.12298

Koerber, S., Sodian, B., Thoermer, C., and Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss J. Psychol.* 64, 141–152. doi: 10.1024/1421-0185.64.3.141

Kuhn, D. (2007). Reasoning about multiple variables: control of variables is not the only challenge. *Sci. Educ.* 91, 710–726. doi: 10.1002/sce.20214

Kuhn, D., and Dean, D. (2004). Metacognition: a bridge between cognitive psychology and educational practice. *Theory Pract.* 43, 268–274. doi: 10.1207/s15430421tip4304

Kuhn, D., and Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychol. Sci.* 16, 866–870. doi: 10.1111/j.1467-9280.2005.01628.x

Kuhn, D., and Pearsall, S. (2000). Developmental origins of scientific thinking. *J. Cogn. Dev.* 1, 113–129. doi: 10.1207/S15327647JCD0101N_11

Kuhn, D., Pease, M., and Wirkala, C. (2009). Coordinating the effects of multiple variables: a skill fundamental to scientific thinking. *J. Exp. Child Psychol.* 103, 268–284. doi: 10.1016/j.jecp.2009.01.009

NGSS Lead States (2013). *Next Generation Science Standards: For States, by States*. Washington, DC: National Research Council.

Ruffman, T. K., Perner, J., Olson, D. R., and Doherty, M. (1993). Reflecting on scientific thinking: Children's understanding of the hypothesis-evidence relation. *Child Dev.* 64, 1617–1636. doi: 10.1111/j.1467-8624.1993.tb04203.x

Schulz, L. E., and Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Dev. Psychol.* 43, 1045–1050. doi: 10.1037/0012-1649.43.4.1045

Schulz, L. E., and Gopnik, A. (2004). Causal learning across domains. *Dev. Psychol.* 40, 162–176. doi: 10.1037/0012-1649.40.2.162

Schulz, L. E., Gopnik, A., and Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Dev. Sci.* 10, 322–332. doi: 10.1111/j.1467-7687.2007.00587.x

Schulz, L. E., and Sommerville, J. A. (2006). God does not play dice: causal determinism and preschoolers' causal inferences. *Child Dev.* 77, 427–442. doi: 10.1111/j.1467-8624.2006.00880.x

Slaughter, V., and Lyons, M. (2003). Learning about life and death in early childhood. *Cogn. Psychol.* 46, 1–30. doi: 10.1016/S0010-0285(02)00504-2

Sobel, D. M., Erb, C. D., Tassin, T., and Weisberg, D. S. (2017). The development of diagnostic inference about uncertain causes. *J. Cogn. Dev.* 18, 556–576. doi: 10.1080/15248372.2017.1387117

Sobel, D. M., and Kirkham, N. Z. (2006). Blickets and babies: the development of causal reasoning in toddlers and infants. *Dev. Psychol.* 42, 1103–1115. doi: 10.1037/0012-1649.42.6.1103

Sobel, D. M., and Munro, S. E. (2009a). Domain generality and specificity in children's causal inference about ambiguous data. *Dev. Psychol.* 45, 511–524. doi: 10.1037/a0014944

Sobel, D. M., and Sommerville, J. A. (2009b). Rationales in children's causal learning from others' actions. *Cogn. Dev.* 24, 70–79. doi: 10.1016/j.cogdev.2008.08.003

Sobel, D. M., Tenenbaum, J. B., and Gopnik, A. (2004). Children's causal inferences from indirect evidence: backwards blocking and Bayesian reasoning in preschoolers. *Cogn. Sci.* 28, 303–333. doi: 10.1016/j.cogsci.2003.11.001

Sodian, B., Zaitchik, D., and Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Dev.* 62, 753–766. doi: 10.2307/1131175

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., and Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychol. Rev.* 115, 155–185. doi: 10.1037/0033-295X.115.1.155

Ullman, T. D., Goodman, N. D., and Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cogn. Dev.* 27, 455–480. doi: 10.1016/j.cogdev.2012.07.005

Zeger, S. L., and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42:121. doi: 10.2307/2531248

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 44:1049. doi: 10.2307/2531734

Zimmerman, C. (2000). The development of scientific reasoning skills. *Dev. Rev.* 20, 99–149. doi: 10.1006/drev.1999.0497

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Weisberg, Choi and Sobel. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 16 September 2020
doi: 10.3389/fpsyg.2020.537672

Development of Conceptual Flexibility in Intuitive Biology: Effects of Environment and Experience

Nicole Betz and John D. Coley**

Department of Psychology, Northeastern University, Boston, MA, United States

Living things can be classified in many ways, such as taxonomic similarity (lions and lynx), or shared ecological habitat (ducks and turtles). The present studies used card-sorting and triad tasks to explore developmental and experiential changes in *conceptual flexibility*–the ability to switch between taxonomic and ecological construals of living things–as well as two processes underlying conceptual flexibility: *salience* (i.e., the ease with which relations come to mind outside of contextual influences) and *availability* (i.e., the presence of relations in one's mental space) of taxonomic and ecological relations. We were also interested in the extent to which salience and availability of taxonomic and ecological relations predicted inductive inferences. Participants were 452 six to ten-year-olds from urban, suburban, and rural communities in New England. Across two studies, taxonomic relations were overwhelmingly more salient than ecological relations, although salience of ecological relations was higher among children from rural environments (Study 1) and those who engaged in unstructured exploration of nature (Study 2). Availability of ecological relations, as well as conceptual flexibility, increased with age, and was higher among children living in more rural environments. Notably, salience, but not availability, of ecological relations predicted ecological inferences. These findings suggest that taxonomic categories (i.e., groups that share both perceptual similarities and rich underlying structure) are a salient way to organize intuitive biological knowledge and that, critically, environmental richness and relevant experience contribute to the salience and availability of ecological knowledge, and thereby, conceptual flexibility in biological thinking. More generally, they highlight important linkages between domain-specific knowledge and domain-general cognitive abilities.

Keywords: intuitive biology, ecological knowledge, conceptual flexibility, experience, salience

OPEN ACCESS

Edited by:
Stella Vosniadou,
Flinders University, Australia

Reviewed by:
Susan Annetta Graham,
University of Calgary, Canada
Andrzej Tarlowski,
University of Economics and Human Sciences in Warsaw, Poland

****Correspondence:***
Nicole Betz
betz.nicolejoy@gmail.com
John D. Coley
j.coley@northeastern.edu

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *24 February 2020*
Accepted: *26 August 2020*
Published: *16 September 2020*

Citation:
Betz N and Coley JD (2020) Development of Conceptual Flexibility in Intuitive Biology: Effects of Environment and Experience. Front. Psychol. 11:537672. doi: 10.3389/fpsyg.2020.537672

INTRODUCTION

The biological world is complex and multidimensional; living things can be interrelated—and therefore organized—in many different ways. For instance, *taxonomic relations* group living things into kinds based on morphological similarities, shared intrinsic features, or common membership in superordinate categories. Orthogonal to relationships based on intrinsic features, a large part of

our knowledge about the biological world relates to extrinsically or thematically shared features. For example, organizing living things by *thematic* relations can describe ecological relationships such as a common habitat, or interactions such as predation. Both taxonomic and ecological relations, although far from exhausting potential relations among living things, are useful classifications, and both emphasize some types of commonalities at the cost of others.

Given that there are multiple ways to meaningfully organize plants and animals, what factors influence the relations that we use to group organisms or to make inductive generalizations about them? In the present studies, we used a card sorting task to explore developmental changes in the relative salience and availability of taxonomic and ecological relations among living things. We also examine the effects of two overarching types of active experiences—unstructured (e.g., exploring nature) and structured (e.g., going to zoos)—and passive experience (e.g., living in a rural area) thereon. Of particular interest is the development of the ability to organize the biological world through multiple frameworks–an ability likely to be related to both richness of biological experience (Shafto and Coley, 2003; Coley, 2012) and the development of more general executive functions (Diamond, 2013; Zaitchik et al., 2014).

This type of *conceptual flexibility* (Blaye et al., 2006) is critical in learning biological science as well as other STEM fields (Coulson et al., 1997; Hoehn and Finkelstein, 2018). In biology, one must be able to think about systems and organisms at a variety of different levels of analysis –the function of an organic molecule in a cell, the function of a cell in an organ, the function of the organ in a body, the function of a body in an ecosystem. Developing such flexible understanding is an important step toward learning to 'think like a biologist.' Thus, by investigating the developing understanding that organisms can be cross-classified in multiple ways, and the impact of active and passive experience thereof, we examine the development of conceptual systems that form an important basis for learning science. Further, this research evaluates the impact of ecologically valid experiences that impact the degree to which children come into the classroom with an ability to think flexibly about relations among organisms.

Beyond a 'Thematic to Taxonomic Shift'?

For many years, conventional wisdom among cognitive development researchers held that young children's concepts are organized in terms of thematic relations, and that only later in development do they use taxonomic relations (e.g., Inhelder and Piaget, 1961/1999; Bruner et al., 1967; Smiley and Brown, 1979; see also Lin and Murphy, 2001 for a review). Thematic concepts were thought to be more accessible to young children because they are more concrete; the associated objects are perceived together and interact directly in salient ways. As they age, children begin to understand how objects are linked based on deeper or less obvious shared properties, which gives rise to taxonomic relations. Older children and adults were thought to primarily rely on purportedly more abstract and logical taxonomic relations, and less on allegedly primitive thematic relations. Based on this view, conceptual development was characterized as a shift from thematic to taxonomic relations (Smiley and Brown, 1979).

More recently, this "thematic-to-taxonomic" shift has been questioned in at least two ways. First, researchers consistently find that young children readily use taxonomic relations to sort objects and guide inferences (e.g., Scott et al., 1982; Gelman and Markman, 1986; Gelman and Coley, 1990; Waxman and Namy, 1997; Nguyen, 2007). For example, preschoolers expect that novel words (e.g., 'See this *wem*? Find another one) refer to taxonomic rather than thematic relations (Markman and Hutchinson, 1984). Likewise, work on the development of intuitive biology has demonstrated an early preference for taxonomic relations for guiding biological *inductive inferences*, or projecting features believed to be true of one class to another related class (e.g., Coley and Vasilyeva, 2010). For example, children as young as 2.5 years of age rely on taxonomic relations (e.g., *birds*) when projecting potentially shared properties (e.g., *lives in a nest*; Gelman and Coley, 1990). Indeed, even 14–16 month old infants' use of labels suggests sensitivity to the importance of membership in taxonomic categories to guide inferences (Keates and Graham, 2008; see also Ferry et al., 2010; Graham et al., 2016; Switzer and Graham, 2017). This and other work (for reviews, see Coley et al., 2002; Coley and Muratore, 2012) suggest that taxonomic relations based on intrinsic, inherent similarities can be more salient for young children than thematic relations. Together, these and other findings (e.g., Nguyen and Murphy, 2003; Coley et al., 2005; Thibaut et al., 2016) debunk the notions of the inaccessibility of taxonomic relationships at an early age, and a ubiquitous early preference for thematic relations in young children.

It's important to note an ongoing debate among researchers in conceptual development on the nature of early category representations. The work reviewed above suggests that young children represent categories as richly structured conceptual representations, entailing assumptions of shared non-obvious features and inductive potential (e.g., Gelman and Coley, 1991; Gelman, 2003; Noyes and Keil, 2019). In contrast, other researchers characterize the process of category acquisition as a progression from simple perceptual groupings to more complex, abstract concepts (e.g., Sloutsky and Fisher, 2004; Sloutsky, 2010), and that early categories and inductions *primarily* stem from perceptual similarity (Sloutsky et al., 2007; Fisher, 2015). By this account, the early salience of "taxonomic" categories stems from shared perceptual features rather than rich abstract representations. However, for our purposes, this debate is largely tangential. Our primary goal is to contrast the salience of categories based on taxonomic, or intrinsic relations–be they perceptual or abstract–with the salience of categories based on thematic, or extrinsic relations.

In addition to evidence suggesting early salience of taxonomic relations, cognitive scientists also increasingly acknowledge the salience and potency of thematic relations for adults, especially among those with relevant expertise. Indeed, thematic relations can be as salient as taxonomic relations across a broad range of categorization and reasoning tasks. For example, adults often prefer to sort entities based on meaningful thematic relations

(e.g., pairing *sheep* with *wool*) over taxonomic (e.g., pairing *sheep* with *goat*; Lin and Murphy, 2001). Further, thematic categories (e.g., breakfast foods) importantly guide people's representation of food categories (Ross and Murphy, 1999). Moreover, expert categorization of plants and animals are more likely to be dominated by thematic relations than novices' (e.g., López et al., 1997; Shafto and Coley, 2003), suggesting that salience of thematic relations may be characteristic of expert knowledge (Coley et al., 2005; Coley and Betz, 2018). These and other findings (Medin et al., 1997; Lin and Murphy, 2001; De Deyne et al., 2016) highlight the potency and use of thematic relations in refined conceptual systems, and the potential role of experience in augmenting the salience of thematic relations.

Flexibility in Conceptual Relations

Rather than characterizing conceptual development as a shift from reliance on one type of relation to reliance on another, it seems more pertinent to acknowledge that children and adults flexibly use a range of conceptual relations to understand and predict their world (e.g., Deak, 2003; Blaye et al., 2007; Nguyen, 2007; Coley and Vasilyeva, 2010). By this account, individuals have access to multiple types of relations around which they can organize the same sets of entities; taxonomic and thematic relations are some examples of these organizational options. Flexibility in conceptual relations describes how the usefulness and subsequent recruitment of different conceptual relations might shift across contexts based on one's knowledge. For example, both thematic and taxonomic relations can provide a rich inductive basis for different types of biological judgments. While taxonomic relations are more potent for making predictions about shared properties that are transmitted through genes, one type of thematic relation—ecological—is more potent for making predictions about shared properties that can be transmitted through contact, such as contagious diseases (e.g., Shafto and Coley, 2003; Coley and Vasilyeva, 2010; Coley, 2012). From this perspective, development involves an increase in the ability to flexibly use a range of different conceptual relations in a given domain. This ability is dependent on both domain-general executive functions (inhibitory control, cognitive flexibility) and domain-specific knowledge of relevant conceptual relations.

Although multiple relations might come to mind as *available* ways to organize entities, these relations often differ in their *salience* to different individuals, and therefore vary with respect to the likelihood that they will be used in any given situation (Shafto et al., 2007b). By "availability" we mean whether knowledge of particular conceptual relations is represented in semantic memory, or readily discernible given the information that is represented in semantic memory. Given that it is possible to hold multiple representations of relations among concepts in a given domain (e.g., Spiro, 1988), availability refers to which representations a particular individual actually holds. By "salience" we mean the ease with which knowledge of a particular type of conceptual relation comes to mind or is retrieved; different conceptual relations, even if all are available to an individual, can vary in their salience. The independence of these two ideas can be seen in the work of Ross and Murphy (1999) regarding food categories. They found that when given a food like *broccoli*, undergraduate participants readily and spontaneously generated both taxonomic (e.g., *vegetables*) and thematic (e.g., *dinner foods*) categories to which the food belonged, demonstrating that both types of categories are available. However, they also found that priming thematic categories (e.g., *breakfast foods*) increased the perceived similarity of thematic matches (e.g., *bagels* and *orange juice*) whereas priming taxonomic categories (e.g., *meat*) had no effect on the perceived similarity of taxonomic matches (e.g., *bacon* and *beef*), suggesting that taxonomic categories were already highly salient and therefore were unaffected by priming, whereas thematic categories were initially less salient and therefore priming increased their salience. Similarly, among students at an urban university, the ability to make thematic inferences about novel properties of animal species diminished under time pressure, whereas time pressure had no effect on taxonomic inferences (Shafto et al., 2007a), suggesting that although both types of relations were available, taxonomic relations among animals were more salient than thematic (i.e., ecological) relations in this population.

Both the availability and salience of conceptual relations seem to be influenced by knowledge and context (for reviews, see Coley et al., 2005; Shafto et al., 2007b). For example, undergraduates without marine expertise rely on taxonomic relations when making any generalizations about fish; in contrast, commercial fishermen only rely on taxonomic relations when reasoning about unknown novel properties (e.g., *has endo inside*), but used ecological relations to reason about relevant known properties, such as disease (Shafto and Coley, 2003). Therein, experts demonstrate greater conceptual flexibility, presumably due to the availability of a wider range of knowledge about relations among marine species (Shafto and Coley, 2003). This and other work (e.g., López et al., 1997; Medin et al., 1997; Bailenson et al., 2002; Ross et al., 2003; Medin et al., 2006) suggest that experience in nature may influence the flexible ability to reason about biology based on multiple types of relations.

Knowledge and context may also influence the development of conceptual flexibility in the domain of biology. Notably, context could refer not only to task constraints, but also to the information that is consistently present in one's environment, such as the rurality of their home. For example, Coley (2012) examined inductive inferences in a forced-choice property projection task. Children in grades $K-6$ from urban, suburban, and rural communities were taught a novel property (i.e., *disease*, or *insides*) about one species (e.g., *cattail*), and asked which of two target species were more likely to share the property. One target species was related taxonomically but not ecologically (e.g., *dandelion*), whereas the other was related ecologically but not taxonomically (e.g., *frog*). Overall, the tendency to make taxonomic inferences about insides was similar among children across age, environment and experience. In contrast, the tendency to make ecological inferences about *disease* increased with age and was higher among children with direct unstructured experience in nature (e.g., unsupervised exploration) and among children who lived in more rural areas. Moreover,

pronounced patterns of selective inference were evident earlier among children from rural communities. Six-year-olds from rural communities preferred to make ecological inferences about novel diseases and taxonomic inferences about novel insides; these patterns were not evident until later among children from suburban or urban communities. This suggests that children with more opportunities to interact directly with relatively intact ecosystems may demonstrate precocious conceptual flexibility with respect to inductive reasoning about plant and animal species.

If conceptual flexibility in the domain of biology emerges earlier in rural children, it would underline the importance of ambient environment and relevant experiences in the development of knowledge of conceptual relations. However, the evidence from Coley (2012) is indirect; children reasoning about *insides* preferred taxonomic inferences, whereas children reasoning about *disease* preferred ecological inferences, but the flexibility of individual children was not tested. Moreover, it does not pinpoint the source of the differences. There are at least two different ways in which young rural and urban children might differ in this respect: in terms of availability of conceptual relations, and in terms of salience. First, the differences in conceptual flexibility may be attributable to differences in *availability* of ecological relations (Shafto et al., 2007b). Unlike rural children. younger urban children may not understand how to meaningfully connect local organisms ecologically because they lack experience with these habitats. Indeed, modern suburban and urban children are less likely to learn about ecological relations through unstructured, exploratory play because rural children spend relatively more time in nature where they could informally observe which local animals cohabitate (e.g., Bunting and Cousins, 1985; Beach, 2003; Coley, 2012). If this is the case, younger urban children might rarely, if ever, organize living things based on shared ecology, whereas younger rural children might have such relations available. Alternatively, younger urban children may have both taxonomic and ecological relations available to them–i.e., possess the requisite knowledge of both taxonomic and ecological relations–but prefer taxonomic reasoning due to the relative *salience* of such relations over ecological. If this were the case, we would expect younger urban children to *be able to* organize living things based on shared ecology, but to *prefer* taxonomic relations as their initial response. Thus, to further explore conceptual flexibility of biological reasoning in young children, the current research investigates the developmental trajectory of both availability and salience of ecological and taxonomic relations, and the effects of experience and environment thereon. Beyond describing how children develop intuitive biological thinking, findings will enhance understanding of the underlying conceptual processes that contribute to conceptual flexibility.

The Current Research

In the studies presented here, we address two sets of research questions. First, we examine developmental and experience-based changes in the *salience and availability of taxonomic and ecological relations* in children's intuitive biological thinking. Specifically, we ask: To what extent are taxonomic and ecological relations among living things salient and available to children in grades $K-6$? How does the salience and availability of these relations change with development? How might children's environment and experiences with nature influence the salience of taxonomic versus ecological relations? Do salience and availability of taxonomic and ecological relations predict patterns of inductive inference? Second, we probe *conceptual flexibility* in the domain of biology (i.e., the ability to shift between taxonomic and ecological thinking about living things). Specifically, we ask: To what extent are children able to demonstrate conceptual flexibility? How does conceptual flexibility change with development? How might children's environment and experiences with nature influence conceptual flexibility?

To answer these questions, we conducted two studies wherein we sampled 452 children in $K-6$th grade from urban, suburban, and rural communities in New England. In Study 1, we used an unguided card sorting task to examine the relative salience of ecological and taxonomic relations among plants and animals. Children sorted pictures into groups, and explained their groupings. We predicted that ecological relations would be more salient for children from rural communities, who have more opportunities for experience with shared local ecological relations, and would also become more frequent with age.

In Study 2, we expanded on this methodology in several ways. We sampled a much larger group of children from a wide range of communities in Massachusetts. We asked children to perform two sorts of the same stimuli, thus eliciting information on what relations are *available* (second sort) as well as which relations are *salient* (first sort). This allowed us to directly assess conceptual flexibility by observing children's ability to produce both taxonomic and ecological sorts. We also collected data on children's nature-related experiences. This allowed us to investigate the degree to which a relevant feature of environment (i.e., population density of hometown) and direct experience may impact both the salience and availability of ecological and taxonomic relations. We predicted that both passive experience with nature that comes from living in a more rural environment, as well as direct experiences with nature would lead to increased salience and/or availability of ecological relations, and increased conceptual flexibility in the biological domain.

Finally, in order to test the implications of salience and availability of taxonomic and ecological relations for generalization, we examined relations between sorting responses and previously published data on patterns of inductive inference in the same sample of children (Coley, 2012). Of interest is whether the salience and/or availability of ecological relations in sorting would predict an increase in the flexible use of ecological relations when making inductions.

By investigating the development of the understanding that organisms can be cross-classified in multiple ways, and the impact of active and passive experience on this understanding, we are examining the development of conceptual systems that form an important basis for learning science, and evaluating the impact of relevant ecological experiences that would impact the degree to which children come into the classroom with an ability to think flexibly about relations among organisms.

STUDY 1

In Study one, we investigated the relative salience of taxonomic and ecological relationships for 6- through 10-year-olds using an unguided card sorting task, and the effects of experience thereon. A relatively unconstrained way to examine the relative salience of taxonomic versus ecological relations is to present a set of stimuli in which both kinds of relations are present, and note which relations children use to spontaneously sort the stimuli. Thus, participants were asked to sort a set of 15 detailed color drawings of plants and animals that were designed to fall into orthogonal taxonomic and ecological groups. Of interest is the relative frequency of taxonomic versus ecological groupings and explanations in each population of children. If experience leads to increased salience of ecological relations, then the frequency of ecological groups and ecological explanations should be higher for children from rural communities than for children from urban communities.

Method

Participants

One hundred and six children from kindergarten through 6th grade were recruited through elementary schools during the 2000–2001 school year. The vast majority of participants spontaneously sorted all the cards and provided explanations, but three participants were omitted for failure to perform the task (i.e., either did not sort all cards or did not provide explanations), yielding one hundred and three total cases. For our analyses, we classified children as rural or urban based on the population density of their town. Although not a direct measure of potentially relevant experience, children from less densely populated areas likely have increased opportunities to interact with plants and animals in relatively intact ecosystems. Fifty-five participants were from East Boston, which was classified as 'urban' (population density 12,166 people per square mile). Forty-nine participants were from one of two different rural communities: North Berwick, Maine (population density 112 people per square mile) and Plainfield, New Hampshire (population density 43 people per square mile). We also grouped participants by age based on their grade into "6-year-olds" (kindergarten; N_{rural} = 16, N_{urban} = 16), "8-year-olds" (second grade; N_{rural} = 17, N_{urban} = 14) and "10-year-olds" (fifth grade; N_{rural} = 16, N_{urban} = 25). These age groups are interesting for several reasons. Kindergartners are just beginning formal education, and as such are unlikely to have had extended exposure to formal science curricula, so their responses are likely to represent informally acquired folk knowledge. Moreover, between the ages of 6 and 10, important developmental changes occur in biological reasoning (see Carey, 1985).

Materials and Design

Stimuli consisted of fifteen 3 in. × 5 in. (7.6 cm × 12.7 cm) laminated cards. Each card contained a realistic color drawing of a local species. Species were chosen to fill a 5 (taxonomic class: *mammals*, *birds*, *insects*, *trees*, and *plants*) × 3 (habitat: *meadow*, *forest*, and *wetland*) matrix (see **Table 1**). Although we acknowledge that 'trees' does not reflect an evolutionarily coherent group of organisms, it is a highly salient and nearly universal category in folk taxonomy distinction between trees and plants is a highly robust folk taxonomic distinction (e.g., Brown, 1984; Berlin, 1999; Coley and Muratore, 2012). Notably, the included stimuli from taxonomic classes also shared many perceptual similarities; we did not include sets of taxonomic stimuli that featured highly perceptually variable members (e.g., a whale and a fox both as mammals).

TABLE 1 | Organisms depicted in sorting task, Study 1.

	Ecological (habitat) category		
Taxonomic category	**Forest**	**Meadow**	**Wetland**
Mammal	Red squirrel	Woodchuck	Beaver
Bird	Screech owl	Meadowlark	Loon
Insect	Gypsy moth	Firefly	Dragonfly
Tree	White pine	Paper Birch	Black Willow
Plant	Fern	Milkweed	Cattails

Procedure

Pictures were arrayed on a table in a random order, and participants were asked to "Put together the things that go together." After sorting all of the pictures into groups, children were asked to explain why they formed each group. The experimenter noted the cards in each group, and wrote down the child's explanation for each group.

Results

Data Coding

Of particular interest in this study were differences in ecological and taxonomic conceptual structures across age and population density. We assessed the degree to which children's sorts reflected taxonomic and ecological relations in two ways. First, we looked at which organisms each child actually placed in the same group. Second, we also examined the *explanations* that each child provided for the groups they formed. These are explained in more detail below.

Groups

For each child, we noted the number of groups formed. We also quantified the content of children's groupings by first counting the number of taxonomic pairs and ecological pairs, and then calculating percentage based on the total number of pairs within each child's sort, akin to the strategy used by Krieter et al. (2016). A taxonomic pair consisted of two species from the same taxonomic category being grouped together (e.g., *woodchuck-beaver or fern-milkweed*). An ecological pair consisted of two species from the same ecological category being grouped together (e.g., *loon-dragonfly or firefly-paper birch*). This calculation allowed us to determine the extent to which children's groups corresponded to our *a priori* assignment of species into taxonomic and ecological groups.

Explanations

Children's explanations for their groupings were transcribed and coded independently for content by 3–4 trained coders

using a constant comparative method (Glaser and Strauss, 1967; Boeije, 2002). Rare cases of disagreement were resolved via discussion. Explanations for each grouping were initially coded into four categories. *Taxonomic* explanations were based on common membership in named categories (e.g., "bugs," "animals," and "trees"). *Ecological explanations* were based on interactions among organisms or between organisms and their environment (e.g., "squirrels like trees," "water animals," and "birds live in trees"). *Appearance* explanations were based on surface attributes or properties such as size, color, shape, or appendages (e.g., "feathers," "hair," and "these look the same"). Finally, uninterpretable or non-sensical or irrelevant explanations were coded as *Other*. Except for the last one, categories were not mutually exclusive, and responses could fall into multiple categories. Inclusion of explanations allowed us to capture when children's rationales for pairings diverged from our *a priori* predictions. For example, children may have paired screech owl, meadowlark, and loon together, and explained that "they all live in trees." Although this would be coded as three taxonomic pairs, the explanation highlights the observation that they all share an extrinsic relation to trees, which we would consider ecological. Thus, examining explanations as well as groups gives us a fuller picture of children's subjective basis for categorizing these organisms.

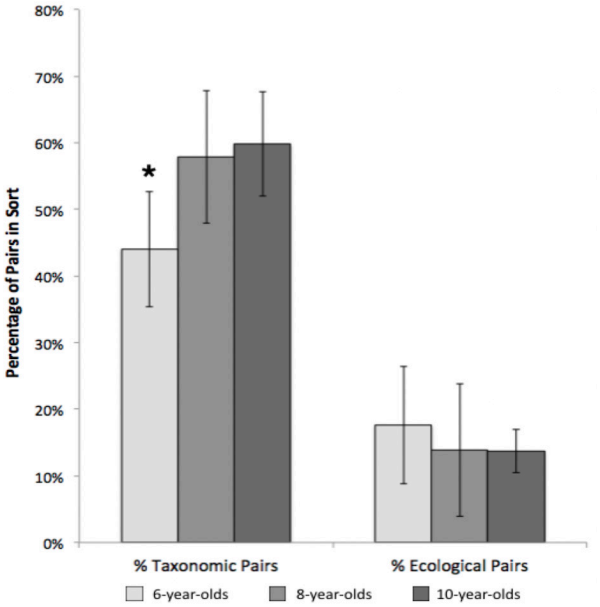


FIGURE 1 | Age differences in percentages of taxonomic and ecological pairs, Study 1. Error bars represent 95% confidence intervals. *$p < 0.001$.

Performance on Sorting Task

Performance on the sorting task was taken as an index of the relative salience of different relations among stimulus species. Of interest are the relative frequency with which children paired ecologically and taxonomically related species, the content of children's explanations for their groupings, and differences in these measures as a function of age and environment.

Groups

On average, children sorted the 15 cards into 4.4 groups (SD = 1.54). To investigate the effects of age and environment on the frequency with which taxonomic and ecological pairs were grouped together, we conducted a 2 (population group: urban, rural) by 3 (age group: 6-, 8-, and 10-year-olds) by 2 (percentage pair: taxonomic, ecological) mixed ANOVA. Overall, sorts were comprised of a higher percentage of taxonomic pairs (M = 55%) than ecological pairs (M = 15%), $F(1,98) = 137.61$, $p < 0.001$, $\text{eta}^2 = 0.58$. We also found that participant age significantly influenced the sorted groups, as demonstrated by an age x pair type interaction ([$F(1, 98) = 3.55$, $p = 0.033$, $\text{eta}^2 = 0.07$)]. Specifically, 6-year-olds created fewer taxonomic pairs than 8-year-olds or 10-year-olds, who did not differ (Tukey HSD $p < 0.05$), whereas age had no effect on number of ecological pairs (see **Figure 1**). We found no significant differences between children from rural or urban communities on number of taxonomic or ecological pairs produced.

Explanations

Each child was given 2 scores, corresponding to the proportion of groups justified with *taxonomic*, or *ecological* explanations. Explanations based on appearance were relatively rare (M = 10%), and did not differ in frequency for urban versus rural children (t = 1.045, p = 0.299). Although explanations corresponded somewhat to pairs, a lack of significant correlation between frequency of pairs and explanations (see **Table 2**) suggests that these two metrics might indeed capture different aspects of children's sorting behaviors.

To investigate the effects of age and environment on children's explanations for their groups, we conducted a 2 (Population Group: Urban and Rural) × 3 (Age Group: 6-, 8-, and 10-year-olds) × 2 (explanation: Taxonomic and Ecological) mixed ANOVA, with population group and age as between-subjects variables. Overall, percentage of taxonomic explanations (M = 47%) were higher than those for ecological (M = 35%), $F(1,98) = 4.09$, $p = 0.046$, $\eta^2 = 0.04$. In contrast to findings for groups, children from urban and rural communities showed different patterns of explanations for their groupings, as evidenced by a population group × explanation interaction, $F(1,98) = 4.16$, $p = 0.04$, $\eta^2 = 0.04$ (see **Figure 2**).

Simple effects analyses revealed that for urban children, taxonomic explanations were more frequent than ecological explanations, $t(54) = 3.04$, $p = 0.004$, Cohen's $d = 0.41$. In contrast, for rural children, taxonomic and ecological explanations were equally frequent. Finally, t-tests directly comparing relative frequencies of each explanation in the two populations revealed that rural and urban children did not differ in the frequency of taxonomic explanations, but rural children produced significantly more ecological explanations than urban children, [$t(102) = 2.55$, $p = 0.012$, $d = 0.50$]. No effects of age on explanations were observed.

Discussion

These results suggest that taxonomic relations are highly salient to children from both urban and rural communities; children

TABLE 2 | Correlations between percentage pairs and explanations, Study 1.

	Explanations	
	Taxonomic	**Ecological**
Pairs		
Taxonomic	0.166^{+}	−0.071
Ecological	−0.058	0.128

Table reports Spearman's correlation coefficients. $^{+}p < 0.10$.

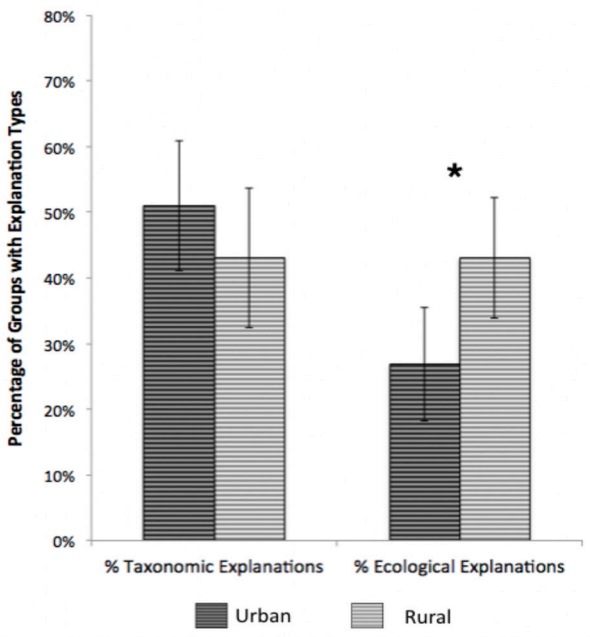


FIGURE 2 | Population group differences in percentages of taxonomic and ecological explanations of sorted groups, Study 1. Error bars represent 95% confidence intervals. $^{*}p < 0.001$.

at all ages showed an overwhelming preference to group together species from the same taxonomic class (e.g., *mammals* and *trees*) rather than from the same ecological habitat (e.g., *wetlands species*). Taxonomic explanations were also relatively frequent in both population groups at all ages; almost half of all groupings were explained on taxonomic grounds. Although there was an increase in taxonomic pairs with development, taxonomic relationships still dominated the 6-year-olds' sorts and explanations. Thus, these results reinforce previous findings showing the importance of taxonomic relations in organizing biological knowledge (e.g., Coley et al., 2002; Coley and Vasilyeva, 2010; Coley and Muratore, 2012) with the caveat that taxonomic relationships become even more salient between the ages of six and eight.

Because the taxonomic stimuli were more perceptually similar than the ecological ones, a potential alternate explanation for the salience of taxonomic relations within our sample is these perceptual similarities. Although it is clear that taxonomic relations were quite salient to children—they not only grouped together taxonomic relations, but also used taxonomic explanations (e.g., "these are birds") to justify their categories instead of explaining groups based on shared appearances—it is possible that the perceptual similarities may have cued the children toward fixating on shared taxonomic relations among the stimuli set. Indeed, the perceptual similarities that occur within taxonomic groups likely contribute to their salience. Importantly, in the real world, categories that share rich internal underlying structure also tend to share perceptual and morphological similarities; bats and whales are the exception, not the rule. Likewise, the taxonomic categories we used here, because they were drawn from reality, included members who shared perceptual as well as deep similarities. As such, although we demonstrate a clear early preference for taxonomic categories, we cannot determine whether this early preference was based on assumptions of shared rich structure, perceptual similarity, or some combination thereof.

Critically, in addition to the clear importance of taxonomic relations, an examination of children's explanations for their groupings showed that, as predicted, ecological relations were more commonly used to justify groupings among children from rural communities than those from urban communities. Indeed, among children from rural communities such explanations were as frequent as taxonomic explanations, suggesting an important influence of ecological thinking and awareness of ecological relations among rural children not seen for urban children. These results are remarkably consistent with results from studies that link informal experience with nature to increased use of ecological relations among species (e.g., Boster and Johnson, 1989; López et al., 1997; Medin et al., 1997; Shafto and Coley, 2003; Ross et al., 2003; Coley, 2012). This difference likely arises from the richer ecological experience of children living in more rural areas, resulting in knowledge of ecological relations which are not perceptually obvious. We investigate this in more detail in Study Two.

STUDY 2

Study one demonstrated that children living in a rural environment were no more likely to form ecological groupings, but were more likely to explain their groupings in ecological terms than children living in an urban environment. This suggests a relationship between environment and salience of ecological relations in children, but leaves unanswered the question of availability of ecological knowledge among children for whom taxonomic relations are salient. In Study 2, we tested a much larger sample of children from a wider range of communities and ages than in Study 1, and sought to replicate and extend the results of Study 1 in several ways.

While the more salient mode of organization (i.e., taxonomic) was the same across rural and urban populations, a single sorting task might underestimate differences in the availability of ecological knowledge across populations. In Study 2, we therefore asked participants to complete two sorts of the same cards, with specific direction to sort the cards in two different ways. While we expected taxonomic relationships to dominate the

first sort again, we hypothesized that performance in the second sort would reflect the availability of ecological relationships after highly salient taxonomic relations were utilized on the first sort. If experience with nature fosters availability of ecological relations, we should find that rural children are more likely to sort species on the basis of ecological relations on a second sort. Moreover, this approach allows us to investigate the development of–and environmental differences in–conceptual flexibility surrounding reasoning about living things, by comparing the basis of children's first and second sorts. Of particular interest was the extent to which children were able to organize the cards based on the two orthogonal dimensions of interest (e.g., Zelazo, 2006; Deák and Wiseheart, 2015).

Further, although the results from Study 1 document a marked difference in explanations for groupings between urban and rural children, they do not pinpoint which aspects of living in a rural environment encourage understanding of ecological relationships. Therefore, in Study 2 we included in-depth activities surveys (inquiring about relevant experiences with plants and animals) and examined the impact of activities in nature on sorting task performance. Specifically, we examine two broad types of potentially relevant experiences: unstructured experiences in nature (e.g., nature exploration, hiking), and learning about plants and animals in structured environments intended to transmit information, albeit informally (e.g., going to zoos or aquariums). In this sample, rural children were more likely to report unstructured activities, but rural, suburban and urban children did not differ in their engagement in structured activities (Coley, 2012). Indeed, children who attend these types of formal educational experiences (i.e., immersive zoo camps) demonstrate marked improvements in their ability to organize animals into taxonomic groups (Unger and Fisher, 2019).

Moreover, we addressed a key methodological concerns with Study 1 by revising our stimuli set. In Study 1, we had a larger proportion of possible taxonomic groups, which could have increased the relative salience of taxonomic relations for our participants. The revised stimuli set in Study 2 therefore contains equal numbers of *a priori* ecological and taxonomic groups.

Additionally, to investigate relations between the distinct but related conceptual processes of intuitive biological categorization and inductive reasoning, we examined whether the relative salience and availability of ecological and taxonomic relations during the sorting task predicted children's tendency to draw inductive inferences based on taxonomic or ecological relations.

Method

All of the data reported in Study 2 were collected as a part of a larger study. The inductive inference results, and their relations to the activities measures, are reported in detail in Coley (2012). Here we focus on the sorting data, which have not been reported elsewhere, and relations between the children's sorting and their activities and inferences.

Participants

Three hundred sixty-two children from kindergarten through sixth grade were recruited through elementary schools and after-school programs from 30 communities in Massachusetts. Data were collected between 2004 and 2007. The vast majority of participants spontaneously sorted all of the cards into groups. Fourteen participants were excluded from analysis for failing to sort all of the cards on either the first or second sort, resulting in a final sample of 348. Sampled communities ranged in size from 9 people per square mile to 15,400 people per square mile, with samples from rural (N = 112), suburban (N = 122), and urban (N = 118) communities. We also sampled children ranging from 4.5 to 12.75 years of age, including samples from grades $K-1$ (N = 117), grades 2–3 (N = 125) and grades 4–6 (N = 106). Children were roughly 60% girls and 40% boys; this distribution held in each subgroup.

Materials, Design, and Procedure

Sorting task

Stimuli consisted of nine 3 in. × 5 in. (7.6 cm × 12.7 cm) laminated cards. Each card contained a realistic color drawing of a local species. Species were chosen to fill a 3 (taxonomic class: *bird*, *invertebrate*, and *plant*) × 3 (habitat: *meadow*, *forest*, and *wetland*) matrix. Items are described in **Table 3**. We cut down on the number of cards included in this task for two reasons. First, a smaller number of cards would be less of a burden for the children to sort twice. Second, fewer taxonomic relationships allowed us to have the same number of possible taxonomic and ecological pairs, in case the presence of additional taxonomic groups in the sorting array increased the salience of taxonomic over ecological groups. As in Study 1, the stimuli within taxonomic categories did share more perceptual similarities than the stimuli within ecological categories.

Children were interviewed individually in a quiet area at their school or after-school program. The pictures were arrayed on a table in a random order, and participants were instructed to "Put together the things that go together." After they sorted the cards, they were asked to explain why they had made each group. Then the cards were shuffled and spread out again and the children were asked if they could put the cards together in a new way, and then to explain their new groupings. The experimenter noted the groups for each of the two sorts, and transcribed children's explanations.

Triad task

Stimuli consisted of 32 8.5 in. × 11 in. laminated cards. Each card featured three realistic color drawings of plants or animals: one premise species and two target species. The premise species was alone at the top of the card and the two target species were equidistant from the premise and located at the bottom of the card. One target was taxonomically related to the premise,

TABLE 3 | Pictures used in sorting task, Study 2.

	Ecological category		
Taxonomic category	**Forest**	**Meadow**	**Wetland**
Bird	Woodpecker	Blue jay	Duck
Invertebrate	Termite	Grasshopper	Slug
Plant	Pine tree	Milkweed	Cattails

but ecologically unrelated while the other target was ecologically related to the premise, but taxonomically unrelated. See **Table 4** for a list of sample triad stimuli, and see Coley (2012) for full details on the triad task.

Participants were told that a blank property was true of the premise category and asked which of the two target categories would share that property. Half of the participants were told that the target category "has stuff inside called *andro*." This was meant to represent an internal property that we predicted would be understood to be anatomical. The other half of participants were told that the target category "has a disease called *andro*." Blank property names changed across each trial.

Activities survey

Prior to the experimental task, each child was given a brief survey about their activities. Within the survey, children were asked questions about their hobbies and interests, with follow-up questions and encouragement to solicit additional information. The survey had questions about hobbies and activities that could have implications for intuitive biology, including pet ownership, gardening, zoos, parks, aquariums, camping, hiking, hunting, and fishing.

Results

Data Coding

Sorting task

Each child's performance on both Sort 1 and Sort 2 yielded two scores: first, the percentage of taxonomic pairs (e.g., woodpecker + blue jay) and ecological pairs (e.g., duck + cattails) out of all pairs produced by the child, and second, the percentage of groups justified by taxonomic and ecological relations (coded as in Study 1). As in Study 1, the percentage of pairs reflected the children's adherence to our predicted taxonomic and ecological categories, while the percentage of justifications reflected the children's perceptions of their groupings. Percentage of taxonomic and ecological pairs were calculated in the same way as in Study 1. However, we used a more comprehensive coding approach for explanations in Study 2. In general, performance on the initial sort was taken as an index of the relative salience of different relations among stimulus species. Performance on the second sort was taken as an index of the availability of children's knowledge of relations among species after the initially salient relations had been identified.

Children's explanations were again transcribed and coded independently for content by 3–4 trained coders using a constant comparative method (Glaser and Strauss, 1967; Boeije, 2002). Rare cases of disagreement were resolved via discussion. Explanations for each grouping were initially coded into the following categories, adapted from Coley and Vasilyeva (2010: (t) taxonomy, (e) general interaction, (e) diet interaction, (e) habitat interaction, (e) similar habitat, (e) similar diet, (n) similar behavior, (n) similar appearance, (n) general similarity, and (n) other. These original codes were thereafter reduced into three dimensions: taxonomic (t), ecological (e), and neither (n). See **Supplementary Table 1** for detailed explanations of dimension reduction, example codes, and prevalence of each subtype of explanation.

TABLE 4 | Sample items used in induction task, Study 2.

Premise category	Taxonomic match	Ecological match
Oak tree	Grass	Squirrel
Cattail	Dandelion	Beaver
Herring	Clownfish	Penguin

Conceptual flexibility

We also created two indices of conceptual flexibility that assessed changes in pairs and in explanations across the two sorts. For pairs, we computed the absolute difference in the percentage of ecological pairs between sort one and sort two, and in the percentage of taxonomic pairs between sort one and sort two. We then averaged the two differences to yield an index of the degree to which responses on the two sorts differed. We also computed an analogous score for changes in ecological and taxonomic explanations. In both cases, higher values represent larger changes in the relative use of taxonomic or ecological relations from sort 1 to sort 2, which we take as evidence of conceptual flexibility. Although there are many ways that flexibility might be quantified, we used this approach to capture the degree to which children spontaneously switched the organizing framework of their sorts from one conceptual dimension of interest to the other.

Activities survey

Interpretable responses to all activities questions were obtained from 252 participants; these were fairly uniformly distributed between urban (N = 82), suburban (N = 72), and rural (N = 98) participants. Children's responses to the activities survey were coded into 23 categories (14 developed to capture open-ended responses, and another nine activities that were specifically queried). These in turn were reduced to 10 activity categories via factor analysis (*Aquarium and Parks*, *Pets and Houseplants*, *Reading and Writing*, *Hunting and Hiking*, *TV and Toys*, *Sports*, *Outdoors and Fishing*, *Video and Board Games*, *Exploring Nature*, and *Animals and Zoos*). For details, see Coley (2012). For purposes of this study, we focused on activity types that were of specific interest to informal biological education: Unstructured Activities (indexed by averaging scores on the *hunting and hiking*, *outdoors and fishing*, and *exploring nature* factors) and Structured Activities (indexed by averaging scores on *aquarium and parks* and *animals and zoos* factors).

Salience and Availability of Taxonomic and Ecological Relations

In this section we investigate the salience and availability of taxonomic and ecological relations by examining children's groupings, and explanations for those groupings, on both the first sort (which we take to index *salient* relations) and the second sort (reflecting knowledge of relations that may be *available* but not highly salient). We also explore the degree to which salience are availability of taxonomic and ecological relations are predicted by activities and environment, and how the salience and availability of these relations predict children's inductive inferences.

What kinds of groupings did children form?

To examine the content of children's groupings, we compared the relative frequency of taxonomic and ecological pairs for the first and second sort. On average, children sorted the nine cards into 3.5 groups on the first sort (SD = 0.82), and taxonomic pairs (M = 76%) represented a higher percentage of the sorts than ecological (M = 11%) pairs, [$t(346) = 23.01$, $p < 0.001$, $d = 1.24$, see **Figure 3**]. Children formed more groups on the second sort (M = 3.6, SD = 1.06) than on the first sort, although the effect was small [$t(347) = 3.12$, $p = 0.002$, $d = 0.17$]. In contrast to the first sort, taxonomic (M = 33%) and ecological pairs (M = 31%) were equally common on the second sort, [$t(334) = 0.78$, $p = 0.44$, $d = 0.04$, see **Figure 3**]. Proportion of taxonomic pairs declined across the two sorts [$t(344) = 16.69$, $p < 0.001$, $d = 0.90$], while the proportion of ecological pairs increased [$t(343) = 10.89$, $p < 0.001$, $d = 0.59$]. This suggests that taxonomic relations were highly salient to most children, but some switched to more ecologically based groupings on their second sort, indicating that ecological relations were nevertheless available.

What factors predicted children's groupings?

To examine which factors predicted children's groupings, we conducted multiple regression analyses utilizing age (in days), population density (people per square mile in child's town), and reported involvement in structured and unstructured activities to predict the number of taxonomic and ecological pairs separately for sort 1 and sort 2. Because of the large range of population density values in our sample (22–15,400), we log-transformed these values. Results of these analyses are summarized in **Table 5**.

First sort. For taxonomic pairs, the overall regression approached significance, but the only significant predictor of performance was age, with older children's sorts featuring higher proportions of taxonomic pairs. For ecological pairs, the overall regression was not significant, although age again significantly predicted performance. In contrast with taxonomic pairs, younger children's sorts had higher proportions of ecological pairs. This suggests that the salience of taxonomic relations increased with age, and the salience of ecological relations decreased with age.

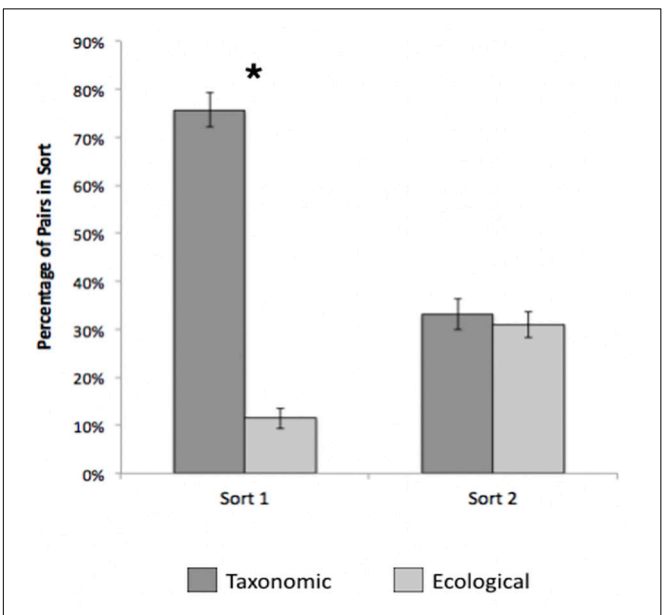


FIGURE 3 | Percentage of taxonomic and ecological pairs across sorts 1 and 2, Study 2. Error bars represent 95% confidence intervals. *$p < 0.001$.

TABLE 5 | Factors predicting groupings on sorts 1 and 2.

	Sort 1		Sort 2	
	Taxonomic pairs	**Ecological pairs**	**Taxonomic pairs**	**Ecological pairs**
R^2	0.038^{+}	0.021	0.018	0.023
Age	0.198**	−0.139*	0.010	0.064
Population density	−0.016	−0.041	0.075	−0.143*
Structured activities	−0.043	0.035	−0.032	−0.001
Unstructured activities	−0.067	0.025	0.122^{+}	−0.047

*Values for age, population density, structured activities, and unstructured activities represent standardized regression coefficients. **$p < 0.01$, *$p < 0.05$, $^{+}p < 0.10$.*

Second sort. For taxonomic pairs on the second sort, the overall regression did not explain significant variance, and no predictors explained significant variance. Nor was the overall regression significant for ecological pairs, although population density explained significant variance. Specifically, proportion of ecological pairs on the second sort decreased with population density (i.e., ecological pairs were more common for children from more rural areas), suggesting that ecological relations became increasingly available for children living in more rural environments.

How did children explain their sorts?

Explanations for groupings on the first sort, like the pairs, were overwhelmingly taxonomic (see **Figure 4**). Paired-samples t-tests revealed that taxonomic explanations (M = 59%) were significantly more common than ecological (M = 19%); $t(347) = 11.476$, $p < 0.001$, $d = 0.62$. In contrast, ecological explanations (M = 55%) were more common than taxonomic (M = 17%) on the second sort; $t(346) = 12.19$, $p < 0.001$, $d = 0.58$. When comparing explanations across the two sorts, the percentage of taxonomic explanations significantly declined [$t(346) = 16.38$, $p < 0.001$, $d = 0.88$], and the percentage of ecological explanations significantly increased [$t(343) = 10.89$, $p < 0.001$, $d = 0.59$]. This is consistent with the findings for pairs presented above; children tended to explain their initial groupings based on taxonomic relations, but explained their second sorting based on ecological relations, suggesting that taxonomic relations were highly salient, but ecological relations were nevertheless available.

What factors predicted children's explanations?

To examine which factors predicted children's explanations, we again conducted multiple regression analyses utilizing age (in days), population density (people per square mile in town, log transformed), and scores on structured and unstructured experience to predict the relative frequency of taxonomic and

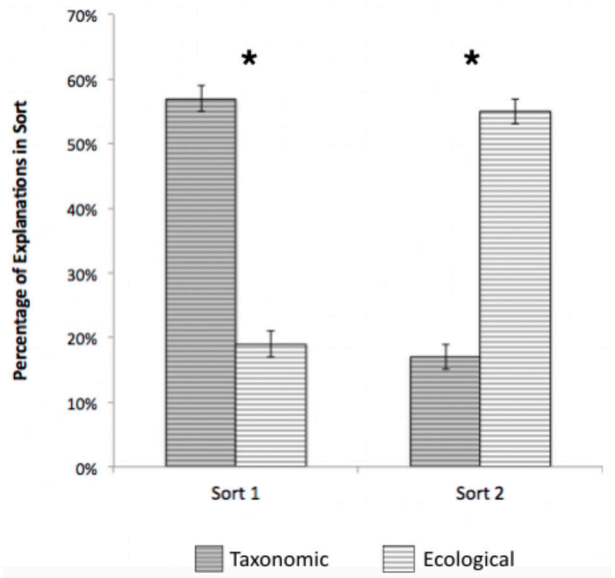


FIGURE 4 | Percentage of taxonomic and ecological explanations across sorts 1 and 2, Study 2. Error bars represent 95% confidence intervals. *$p < 0.001$.

TABLE 6 | Factors predicting explanations on Sorts 1 and 2.

	Sort 1		Sort 2	
	Taxonomic explanations	**Ecological explanations**	**Taxonomic explanations**	**Ecological explanations**
R^2	0.135***	0.032	0.034+	0.060**
Age	0.361***	−0.066	0.096	0.154*
Population density	−0.098	−0.042	−0.121+	−0.153*
Structured experience	−0.038	0.083	−0.102	0.090
Unstructured experience	−0.057	0.147*	−0.001	−0.123+

*Values for age, population density, structured activities, and unstructured activities represent standardized regression coefficients. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, +$p < 0.10$.*

TABLE 7 | Correlations between percentage pairs and explanations, Study 2.

	Explanations, sort 1			Explanations, sort 2	
	Taxonomic	**Ecological**		**Taxonomic**	**Ecological**
Pairs, Sort 1			**Pairs Sort 2**		
Taxonomic	0.720***	−0.730***	Taxonomic	0.472**	−0.542**
Ecological	−0.656***	0.698***	Ecological	−0.336***	0.480***

*Table reports Spearman's correlation coefficients. ***$p < 0.001$, **$p < 0.01$.*

ecological explanations for sort 1 and sort 2. Results of these analyses are summarized in **Table 6**.

First sort. For taxonomic explanations, the overall regression explained significant variance; the proportion of taxonomic explanations increased with age, suggesting that the salience of taxonomic relations increased with age but was unrelated to activities. For ecological explanations, the overall regression did not explain significant variance, but the proportion of ecological explanations increased with unstructured nature activities, suggesting that ecological relations were increasingly salient among children who engaged in one type of relevant experience—unstructured nature exploration.

Second sort. For taxonomic explanations, the overall regression did not reach significance and no factors predicted significant variance. For ecological explanations, the regression was significant; ecological explanations increased with age, and decreased with population density (i.e., were more common among children from more rural communities), suggesting that these factors impacted the availability if not the salience of ecological relations.

The relationship between explanations and groups

As in Study 1, the explanations and groups provided two different types of metrics of performance—adherence to our *a priori* taxonomic and ecological relations versus children's reliance on taxonomic and ecological relations. In contrast to Study 1, we find strong positive correlations between taxonomic and ecological pairs and explanations within each of the sorts (see **Table 7**), a testament to the efficacy of our updated coding system and stimuli. The explanations and pairs in Study 2 therefore seem to converge quite highly. Further, we see a repeated negative relationship between taxonomic and ecological, suggesting that children are emphasizing one type of relation at the expense of the other.

How is performance on the sorting task related to children's inductive inferences?

To examine relations between children's sorting and their inductive inferences, we ran separate regressions using children's first and second sort taxonomic and ecological pairs and explanations to predict their proportion of ecological inferences on the triad induction task (i.e., the percentage of trials for which a child projected the novel property to the ecological target rather than the taxonomic target). Each measure of performance was entered into a separate regression, along with age and log population density as control variables (both of which have been linked to ecological inferences, Coley, 2012). Results are reported in **Table 8**, and see **Supplementary Tables 2, 3** for results from specific properties within the triad task.

Regressions revealed several significant predictors of ecological inferences. Notably, salience of ecological relations, in terms of proportion of both ecological pairs and ecological explanations on sort 1, predicted frequency of ecological inductive inferences. Further, the proportion of taxonomic pairs in sort 1 negatively predicted frequency ecological inferences. No index of performance on sort 2 was related to ecological inferences. Overall, these findings suggest that the tendency to base inferences on ecological relations in the induction task was closely tied to the salience of ecological relations in the sorting task.

TABLE 8 | Relations between indices of performance on sorting task and preferences for ecological inferences in the triad induction task.

Sort	Relation type	Performance index	Standardized regression coefficient
1	Ecological	% Pairs	0.137*
		% Explanations	0.116*
	Taxonomic	% Pairs	−0.138*
		% Explanations	−0.063
2	Ecological	% Pairs	−0.011
		% Explanations	0.058
	Taxonomic	% Pairs	0.085
		% Explanations	0.029

*$p < 0.05$.

TABLE 9 | Factors predicting conceptual flexibility scores for pairs and explanations.

	Flexibility index	
	Pairs	Explanations
R^2	0.123***	0.144***
Age	0.313***	0.373***
Population density	−0.131*	−0.089
Structured activities	−0.143*	−0.006
Unstructured activities	−0.065	−0.043

*Values for age, population density, structured activities, and unstructured activities represent standardized regression coefficients. ***$p < 0.001$, *$p < 0.05$.*

What factors predicted conceptual flexibility?

In this study, conceptual flexibility reflects the extent to which children changed between taxonomic and ecological organizational frameworks across their two sorts. In other words, we were primarily interested in the degree to which children produced different kinds of sorts—i.e., changed the underlying dimension along which they sorted organisms—rather than simply producing different groups. We therefore quantified conceptual flexibility as change (regardless of direction) in the percentage of taxonomic and ecological pairs, and change in the percentage of taxonomic and ecological explanations, between sort 1 and sort 2. Higher scores reflect more change from the first to second sort, and hence more flexibility. The two sorts provide qualitatively different information, with the first sort assessing the most salient relations and the second sort assessing additional relations that are available after the most salient ones were taken away. To the extent that children tended to rely on one underlying dimension for each sort, our approach could quantify the degree to which children were able to change this basis of their responses across sorts. This analysis was justified based on preliminary assessments of the data that revealed that sorts most often were dominated by either taxonomic or ecological relations, rather than split in between both types of relations (see **Supplementary Figures 1–4**; this effect is also highlighted by the negative correlations between ecological and taxonomic pairs and explanations presented in **Table 7**).

To examine which factors predicted flexible use of conceptual relations, we computed two multiple regression analyses utilizing age (in days), log population density (people per square mile in town), and structured and unstructured activity scores to predict conceptual flexibility with respect to pairs and explanations, respectively. Results are summarized in **Table 9**.

The regression models explained significant variance in children's conceptual flexibility with respect to both pairs and explanations. Conceptual flexibility regarding pairs increased with age and decreased with population density (i.e., children from more rural environments showed more flexibility). Surprisingly, children who reported more structured activities like aquariums and zoos also showed less conceptual flexibility. Conceptual flexibility regarding explanations increased with age, but was unrelated to environment or experience.

Discussion

Results from the initial sort in Study 2 replicated the findings from Study 1 with a different population and stimulus set. Participants were much more likely to pair species on the basis of taxonomic than ecological relatedness; this tendency increased with age and held constant regardless of children's environment. As in Study 1, taxonomically related stimuli were also perceptual more similar; as such, we cannot determine whether this early preference was based on assumptions of shared rich structure, perceptual similarity, or both.

Further, we again found some evidence of an effect of experience on explanations for sorts, if not sorting *per se*. In Study 1, ecological explanations were more frequent among children from rural environments, whereas in Study 2, ecological explanations increased with unstructured experiences in nature. Moreover, Study 2 revealed that knowledge of ecological relations, if not overly *salient*, is nevertheless *available* to many children. Ecological pairs were as frequent as taxonomic pairs in sort two, and ecological explanations dominated the second sort. Interestingly, both older children and children of all ages from more rural environments were more likely to have ecological relations available to them. These findings suggest two ways in which ecological relations can be learned—both through ambient environment (i.e., living in a more rural area) and through experiences accrued with age.

We examined conceptual flexibility surrounding living things by quantifying change in pairs and explanations from Sort 1 to Sort 2. Results show that conceptual flexibility increased markedly with age, in line with other work (e.g., Blaye et al., 2006; Coni et al., 2019). Notably, living in a more rural environment predicted conceptual flexibility *independently* of age. Thus, children who live in an environment with the potential for environmental exploration not only have ecological relations more available to them, but also were able to flexibly organize plants and animals according to these relations. Surprisingly, results also revealed that children who reported more frequent structured biologically relevant activities, such as visiting aquariums and zoos, demonstrated less flexibility. These patterns may suggest that children who are taught about biology in more structured ways may become entrenched in these taught views and thereby may become less likely to flexibly switch between different types of relations

(i.e., Bonawitz et al., 2009; Dane, 2010). This possibility is discussed further below.

Finally, we demonstrated that the salience, but not availability, of ecological relations predicted ecological inferences. Overall, these findings corroborate Shafto et al. (2007b), demonstrating that salience of relations, rather than mere availability, impacts induction. Such findings emphasize the importance of distinguishing between salience and availability, and begin to describe how the ease with which relations come to mind may impact use of concepts in everyday life.

GENERAL DISCUSSION

From a young age, children simplify the complexities of the biological world by constructing intuitive concepts that organize knowledge and guide predictions about plants and animals. While early research on such conceptual development posited a thematic to taxonomic shift, more recent work instead highlights the importance of viewing development as increasing conceptual flexibility. To investigate this, we used unguided card sorting tasks that were designed with two orthogonal dimensions: taxonomic categories (e.g., birds) and ecological relationships (e.g., pond animals). We addressed two sets of research questions. First, we examine developmental changes in the *salience and availability of taxonomic and ecological relations* in children's intuitive biological thinking. Second, we probe developmental changes in *conceptual flexibility* in the domain of biology (i.e., the ability to shift between taxonomic and ecological thinking about living things). In both cases, we also assessed the impact of environment and activities.

Salience and Availability of Taxonomic and Ecological Relations

Our first set of questions focused on developmental changes in the *salience and availability of taxonomic and ecological relations* in children's intuitive biological thinking. Below we consider the implications of our findings for understanding salience, availability, and relations to inductive inference.

Salience of Taxonomic and Ecological Relations

We take *salience* to refer to the ease with which knowledge of a particular type of conceptual relation comes to mind or is retrieved; different conceptual relations, even if all are available to an individual, can vary in their salience. In these studies, we took children's groupings and explanations on the first performance of a sorting task as an index of the salience of taxonomic and ecological relations; to the degree that such relations readily come to mind upon encountering representations of organisms, they should be used to sort and explain. Across age and experience, taxonomic relations were overwhelmingly more salient than ecological relations. In Study 1 and the first sort of Study 2, children were much more likely to pair animals from the same taxonomic category than animals from the same ecosystem, and were much more likely to explain their groupings in taxonomic than in ecological terms. The salience of taxonomic categories was already evident among the youngest participants in our study, and increased with age. This fits with work stressing the importance of richly structured taxonomic categories for children's understanding of the biological world from early in development (e.g., Scott et al., 1982; Gelman and Markman, 1986; Gelman and Coley, 1990, 1991; Waxman and Namy, 1997; Nguyen, 2007).

As argued above, categories that share rich internal underlying structure also tend to share perceptual and morphological similarities. Likewise, the taxonomic categories we used in the studies presented here, because they were drawn from reality, included members who shared perceptual as well as deep similarities. As such, although we demonstrate a clear early preference for taxonomic categories, our design does not permit us to adjudicate whether this early preference, or indeed any "taxonomic" responses, were based on assumptions of shared rich structure or on perceptual similarity. Although perceptual similarity is nearly always an important guide to richly structured categories, we think there are compelling theoretical (e.g., Murphy and Medin, 1985; Medin, 1989; Ahn et al., 2001) and empirical (e.g., Barsalou, 1983; Gelman and Markman, 1986, 1987; Keil, 1989; Rips, 1989; Heit and Hayes, 2005; Noyes and Keil, 2020) arguments against a strong perceptual view of early category representations. However, the primary aim of the design employed in these studies was to contrast the salience of categories based on intrinsic taxonomic relations–be they perceptual or abstract–with the salience of categories based on extrinsic ecological relations. As such, our results are agnostic with respect to the perceptual-conceptual discussion.

Notably, despite the general salience of taxonomic categories, we also consistently found that the salience of ecological relations was related to environment and experience. In Study 1, although children from rural environments formed as many taxonomic groups as children from urban environments, they were more likely to explain their grouping in ecological terms. In Study 2, children's environment again did not predict groupings, but the proportion of groups explained in ecological terms increased with children's unstructured experiences in nature. Whereas the high salience of taxonomic relations further increases with age, ecological relations are more salient among children who have relevant experiences, particularly those who live in rural communities, and those who reported more unstructured nature exploration. This is consistent with a number of studies showing that both relevant activities (e.g., unstructured exploration) and passive experience (i.e., living in more rural environments) increase access to ecological knowledge (e.g., López et al., 1997; Proffitt et al., 2000; Shafto and Coley, 2003), suggesting that salience of thematic relations may be characteristic of expert knowledge (Coley et al., 2005; Coley and Betz, 2018). These and other findings (Medin et al., 1997; Lin and Murphy, 2001; De Deyne et al., 2016) highlight the potency and use of thematic relations in refined conceptual systems.

Availability of Ecological Relations

By our account, *availability* refers to whether knowledge of particular conceptual relations is represented in semantic memory, or is readily discernible given the information that

is represented in semantic memory. Salient knowledge is necessarily available, but all available knowledge need not (and indeed, cannot) be equally salient. In Study 2, we asked children, after completing their initial sort, to sort the same cards in a different way. We take performance on this second sort as an index of availability of less salient relations; sort 1 allowed children to "discharge" salient knowledge, and sort 2 gave them the opportunity to demonstrate knowledge that, while less salient, was nevertheless available.

Despite being less salient, ecological relations *were* widely available, as indicated by an increase in ecological pairs and an overwhelming preference to provide ecological explanations on sort 2. Notably, we observed a developmental increase in the availability of ecological relations (for explanations). The availability of ecological relations was also consistently related to population density in that–for both pairs and explanations–ecological relations were more available for children from more rural environments. This highlights the potential role of ambient experience in shaping the availability of conceptual relations; in the case of biology, living in an ambient environment with the potential for environmental exploration may encourage the availability of ecological relations.

Do Salience and Availability of Conceptual Relations Predict Patterns of Inductive Inference?

By comparing the sorting performance of children in Study 2 to their performance on a previously published triad induction task (Coley, 2012), we also asked whether the salience and/or availability of taxonomic and ecological relations influence patterns of inductive reasoning. Results clearly showed that children for whom ecological relations were more salient (both pairs and explanations) were also more likely to make ecological inferences (i.e., project a novel property to an ecologically related species rather than a taxonomically related species). In contrast, the mere availability of conceptual relations did not predict inductive preferences. Thus, the ease with which conceptual relations come to mind (i.e., salience), but not the mere presence of such relations in one's conceptual toolbox (i.e., availability) influences the recruitment of such relations when making predictions about the world.

Results also illuminate previous findings on the development of selectivity in inductive inference. Coley (2012) reported that rural children were precociously sensitive to context on the triad induction task described here, with 6-year-old rural participants making more taxonomic inductions when reasoning about *insides*, and more ecological inductions when reasoning about *diseases*. In contrast, urban and suburban children did not show evidence of this context sensitivity until later ages (10 and 8 years, respectively). The current findings shed light on a potential mechanism behind this context sensitivity: heightened baseline salience of ecological knowledge for rural children. Heightened salience of ecological relations might allow rural children to access either taxonomic or ecological knowledge in the context of inductive inference depending on the property being projected. In contrast, ecological knowledge in urban children—even if available—might not be sufficiently salient to allow differential access among younger urban participants.

Salience and Availability as Distinct Constructs

Our results support the idea that the distinction between salience and availability of conceptual relations is a useful one for understanding conceptual development. This distinction is highlighted by performance on the sorting task. There, salience of ecological relations was predicted by children's unstructured activities; in contrast, availability of ecological relations was predicted by rural environment. Thus, active engagement predicted salience whereas living in an environment that provides opportunities for passive nature observation predicted availability. Moreover, salience of relations predicted patterns of inductive inference, whereas availability was unrelated to inferences. Thus, it seems like considering the salience and availability of conceptual relations separately contributes to our understanding of development. These findings build on and expand ideas about availability in inductive reasoning (Shafto et al., 2007b).

Conceptual Flexibility

We posit that to demonstrate conceptual flexibility within a domain, an individual must represent multiple relevant conceptual relations, and also be able to inhibit salient relations and flexibly access and use other available knowledge. As such, it represents both domain specific knowledge and domain-general executive functions. Our findings fit with this dual characterization. We defined conceptual flexibility within our study as the extent to which children changed their reliance on both types of predicted organizational relations (i.e., ecological and taxonomic) across sorts one and two. Not only did we observe strong and consistent increases in conceptual flexibility with age, we also saw increased conceptual flexibility in the domain of biology among children living in more rural areas. This suggests that development of conceptual flexibility is driven by domain-specific experience as well as the emergence of domain-general executive functions.

Results clearly showed a marked increase in conceptual flexibility with development, for both pairs and explanations. This could represent the contribution of development of domain-general executive functions like cognitive flexibility and inhibitory control, and/or the acquisition of taxonomic and ecological knowledge via formal or informal means. Executive functions–typically seen as working memory/information updating, cognitive flexibility/shifting, and inhibitory control (Miyake et al., 2000)–are defined as "adaptive, goal-directed behaviors that enable individuals to override more automatic or established thoughts and responses (Garon et al., 2008, p. 31)." Although basic components are typically in place before children begin formal education (Garon et al., 2008), executive functions like inhibitory control and working memory undergo a gradual developmental trajectory that extend well into late adolescence and young adulthood (Davidson et al., 2006; Diamond, 2013). The increase with age that we observed in conceptual flexibility undoubtedly reflects in part an increase in general executive function being demonstrated in intuitive biological thought. This view fits with that of Spiro (1988), who argued that individuals shift from a single representation to multiple representations,

which acts as a foundation for cognitive flexibility. Similarly, Deak (2003) suggested that individuals first create a range of plausible ways to organize information and that flexibility allows them to select appropriate patterns based on task demands.

The finding of increased conceptual flexibility in the domain of biology among children living in more rural areas suggests an impact of domain-specific experience on the development of conceptual flexibility. Living in rural environments, and having the opportunity to directly observe and interact with species in relatively intact ecosystems, could facilitate the development of conceptual flexibility in a number of ways. Most simply, such experience might provide children living in rural environments with knowledge of ecological relations among species; children living in more urban environments might not have access to such knowledge. In order to demonstrate conceptual flexibility, a child must have multiple conceptual systems available. Perhaps rural children are more flexible because they possess knowledge about ecological relations among living things that they can demonstrate on a second sort, whereas children living in urban environments typically lack such knowledge. One piece of evidence suggesting that this difference cannot be attributed solely to differential knowledge is the negative relationship between the "structured activities" factor and conceptual flexibility. One might expect children who tend to frequent zoos and museums to possess a relatively large reservoir of factual knowledge, but if so, this isn't sufficient to guarantee conceptual flexibility. Another possibility is that more varied experience with living things–such as that afforded by living in a less densely populated locale–facilitates the development of more general executive function capacities. Previously demonstrated linkages between individual differences in the development of theory-like intuitive biological thinking and executive function have been interpreted to show that executive function provides a foundation for the development of intuitive biology (Zaitchik et al., 2014). However, analogous to claims about effects of bilingualism on executive function (e.g., Bialystok and Martin, 2004; Bialystok et al., 2014), perhaps living in a relatively rich biological environment scaffolds the development of executive functions more generally by requiring increased inhibitory control and cognitive flexibility in children as they negotiate their surroundings and construct representations of kinds and relations therein. Our results cannot arbitrate between these two possibilities, but they raise important questions for future research.

Thus, these results support those of Coley (2012), and other work showing an increase in the flexible use of multiple conceptual dimensions among experts (e.g., López et al., 1997; Shafto and Coley, 2003; see Shafto et al., 2007b, for a review). It is important to emphasize that the increase in conceptual flexibility with decreasing population density was specific to intuitive biological thinking. We did not include any general measure of executive function, and moreover, we would hypothesize that any developmental context providing the opportunity for children to observe and interact with complexity would give rise to increased conceptual flexibility in that domain. For example, children growing up in urban settings might show greater conceptual flexibility with respect to social categories, or spatial wayfinding. Nevertheless, these findings raise the possibility that the acquisition of domain specific knowledge or experience could bolster a general cognitive ability and may suggest an important domain-specific component to executive function.

Conceptual Development as Increasingly Flexible Use of Multiple Relations

Our results are more consistent with a view of development as increasing conceptual flexibility than a shift from emphasis on one dimension to another. When looking only at salience of taxonomic and ecological relations, our findings align somewhat with the notion of a thematic to taxonomic shift; taxonomic relations became increasingly salient across development, while *salience* of ecological relations decreased across development. However, counter to the idea that knowledge of taxonomic relations emerges later in development, the responses of even our youngest participants were overwhelmingly taxonomic. Moreover, the availability of ecological relations increased with development and with experience. The early salience of taxonomic relations and increasing availability of ecological relations are inconsistent with the idea that young children prefer thematic relations, and that reliance on them decreases with age. These, along with the fact that conceptual flexibility increased markedly with age, suggest instead that development is more profitably characterized as the increasing emergence of the ability to flexibly utilize knowledge of different relations for different purposes (see also Gelman and Markman, 1986; Waxman and Namy, 1997; Ross and Murphy, 1999; Thibaut et al., 2016; for a review, see Lin and Murphy, 2001).

Implications for Science Education

Informal Experience Facilitates Conceptual Flexibility

Conceptual flexibility is critical in learning biological science as well as other STEM fields. In biology, one must be able to think about biological entities and concepts at different levels of analysis. Indeed, this is a guiding principle of the "Vision and Change" core concepts in undergraduate biology education (American Association for the Advancement of Science, 2011) as well as the "crosscutting concepts" in the *Next Generation Science Standards* for K-12 education (National Research Council, 2012). As such, demonstrating conceptual flexibility might be a signal that a child is beginning to 'think like a scientist' (e.g., Williams et al., 2004). Indeed, formal scientific thought requires the ability to hold and flexibly navigate between multiple different types of relations. Our findings highlight the benefits of informal domain-specific experiences (e.g., Ritter et al., 2012) in developing domain-specific cognitive flexibility. We have shown that simultaneously understanding that organisms can belong to ecological as well as taxonomic groups increases with age, and moreover that the environment in which a child lives explains independent variance in such conceptual flexibility. Thus, children's informal experience impacts the development of conceptual systems that form an important basis for learning science and affects the degree to which children come into the classroom with an ability to think flexibly about relations among organisms.

Informal Experience Can Impede Conceptual Flexibility

While experience can facilitate cognitive flexibility, there also may be a trade-off between the potential outcomes of flexibility and *cognitive entrenchment*—a high level of stability in thinking that can come with expertise or formal education (for a review, see Dane, 2010). Consistent with this idea, we found that frequent participation in more structured (albeit informal) activities, such as visiting aquariums and zoos, predicted less conceptual flexibility. This somewhat counterintuitive relationship between experience and rigidity in thinking aligns with ideas of a 'double-edged sword of pedagogy' (e.g., Bonawitz et al., 2009, 2011) wherein formal instruction not only benefits children's ability to achieve a targeted goal (e.g., successfully making one toy's function work) but also decreases their likelihood to creatively explore (e.g., find other functions of the toy). Structured experiences that are clearly meant to convey information can help children learn more quickly but can also discourage explorative play by suggesting that there are 'right' and 'wrong' ways of thinking about a domain, or that the important information has already been conveyed. To some extent, this could be quite beneficial; it could help individuals converge on more conventional ways of organizing their knowledge, which could facilitate communication and could help individuals focus on relations that are known to have robust inductive utility. At the same time, this could also restrict individuals' creativity and novel interpretations of the world, which might inhibit conceptual flexibility. One factor that could encourage entrenchment at the expense of flexibility might be the lack of engagement with a "dynamic environment within one's expertise domain" (Dane, 2010; p. 589). Together with the theory of pedagogy's 'double-edged sword,' this might elucidate the asymmetry that we found between effects of rural environment (associated with more conceptual flexibility) and structured activities (associated with less conceptual flexibility); dynamic, diverse, and perhaps unguided experiences might help encourage more flexible thinking.

Future Research

These findings raise important avenues for future research both to enhance the field's understanding of conceptual development and to bolster science education. First and foremost, it will be critical to determine how children transition from having multiple relations available to being able to flexibly shift between these relations in real time. To answer this question, it will be useful to describe the interactions between developing domain-specific conceptual flexibility and more domain-general executive functions. While we have speculated that executive functions such as working memory or inhibitory control are important for conceptual flexibility, our results do not directly speak to their interactions. Finding this predicted relationship would have implications for age-related limitations in conceptual development, and would suggest potential interventions to improve conceptual flexibility.

Another important next step involves investigating the role of conceptual flexibility in the biology classroom. Of particular interest is whether increased salience or availability of ecological relations, or increased conceptual flexibility surrounding biological entities, predicts academic success. Such findings would demonstrate the importance of intuitive biology beyond our daily interactions with plants and animals, demonstrating how these informal ways of learning can provide a leg up in school. Likewise, research should also investigate the effects of both structured and unstructured activities in conceptual flexibility. Our findings provide preliminary correlational evidence that both types of activities could factor into a potential trade-off between conceptual flexibility and cognitive entrenchment. Delving deeper into this relationship could help identify a proper amount of structure in activities that provide children with information quickly, but also encourage them to explore and discover some of the answers on their own. This type of activity could not only enhance learning but also create positive attitudes of self-efficacy surrounding knowledge acquisition.

Limitations

As noted above, our study is limited in its ability to explain why children formed taxonomic categories—on the basis of perceptual similarities, due to richly structured representations shared underlying features, or a combination thereof. However, it was never our goal to investigate the basis of taxonomic categorization *per se*, but rather to demonstrate that children are able to flexibly switch between two types of salient categories—taxonomic and ecological—and to elucidate the role of experience in the development of this ability. Therefore, we do not believe that this limitation compromises our conclusions about the development of conceptual flexibility.

Relatedly, our study is limited to some extent by the stimuli that we selected. Some of our taxonomic groups share more functional and perceptual similarities (e.g., the wings of the insects) than others (e.g., the more diversified sets of plants). These stimuli choices could have influenced the types of categories that participants generated. For example, the perceptual similarities in some of the taxonomic categories in our card set could have cued certain patterns of taxonomic groupings that could overestimate the salience of taxonomic relations among sampled children. Further, there are also ecological relations that children may know about that we did not probe within the card sets. This could have led to an underestimation of the salience of ecological relations among the sampled children.

Finally, we only included stimuli featuring local animals within our stimuli set because we expected direct experience with local animals to be relevant to ecological thinking. Thus, we are unable to determine whether the differences we observed between children from rural versus urban environments would generalize to reasoning about unfamiliar or exotic species. On one hand, if ecological relations are more broadly salient to rural children, we might expect rural children to still produce more ecological relations for more exotic species. That is, their experience might increase the salience of ecological relations in general, and therefore transfer to contexts beyond their immediate experience. On the other hand, if the salience of ecological relations specifically stems from experience with the

local species, and perhaps the opportunity to observe those relations directly, this salience may not generalize to more exotic creatures. In this case, we would expect rural and urban children to perform similarity for unfamiliar or exotic species. This question gets at more general issues of knowledge transfer, and is therefore an important avenue for future research.

CONCLUSION

Using an open-ended sorting task, we have shown that taxonomic relations among living things are overwhelmingly salient to children independent of age, environment, and experience. However, by distinguishing between the salience and the availability of conceptual relations, we also show that despite this general salience of taxonomic relations, the availability of ecological relations increased with age and by both direct experience with nature (i.e., through informal exploration) and the population density of their hometown. Critically, we showed that conceptual flexibility–the ability to switch between thinking about organisms taxonomically and ecologically–increases with age but is also predicted by experience. We argue that these results point to an account of conceptual development as the emergence of flexible multidimensional conceptual systems for thinking about domains of experience, as exemplified here in the domain of biology. Finally, our results document the importance of taking children's background into account in science teaching and learning, by demonstrating that life experiences—what children bring to the science classroom—can impact both domain specific and domain general cognitive processes critical for the development of scientific thought.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Northeastern University Institutional Review Board. Written informed consent to participate in this study was provided by the participants' legal guardian.

AUTHOR CONTRIBUTIONS

JC designed the study, collected the data, and assisted with data analysis and writing of the manuscript. NB led the data analysis and writing of the manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Science Foundation under grant no. BCS-0236338 to JC.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.537672/full#supplementary-material

REFERENCES

Ahn, W.-K., Kalish, C., Gelman, S. A., Medin, D. L., Luhmann, C., Atran, S., et al. (2001). Why essences are essential in the psychology of concepts. *Cognition* 82, 59–69. doi: 10.1016/s0010-0277(01)00145-7

American Association for the Advancement of Science (2011). *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washington, DC: American Association for the Advancement of Science.

Bailenson, J. N., Shum, M. S., Atran, S., Medin, D. L., and Coley, J. D. (2002). A bird's eye view: biological categorization and reasoning within and across cultures. *Cognition* 84, 1–53. doi: 10.1016/s0010-0277(02)00011-2

Barsalou, L. W. (1983). Ad hoc categories. *Mem. Cogn.* 11, 211–227. doi: 10.3758/bf03196968

Beach, B. A. (2003). Rural children's play in the natural environment. *Play Cult. Stud.* 5, 183–194.

Berlin, B. (1999). "How a folkbotanical system can be both natural and comprehensive: one Maya Indian's view of the plant world," in *Folkbiology*, eds D. L. Medin and S. Atran (Cambridge, MA: MIT Press), 71–89.

Bialystok, E., and Martin, M. M. (2004). Attention and inhibition in bilingual children: evidence from the dimensional change card sort task. *Dev. Sci.* 7, 325–339. doi: 10.1111/j.1467-7687.2004.00351.x

Bialystok, E., Poarch, G., Luo, L., and Craik, F. I. (2014). Effects of bilingualism and aging on executive function and working memory. *Psychol. Aging* 29:696. doi: 10.1037/a0037254

Blaye, A., Bernard-Peyron, V., Paour, J. L., and Bonthoux, F. (2006). Categorical flexibility in children: distinguishing response flexibility from conceptual flexibility; the protracted development of taxonomic representations. *Eur. J. Dev. Psychol.* 3, 163–188. doi: 10.1080/17405620500412267

Blaye, A., Chevalier, N., and Paour, J. L. (2007). The development of intentional control of categorization behaviour: a study of children's relational flexibility. *Cogn. Creier Comport.* 11, 791–808.

Boeije, H. (2002). A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Qual. Quant.* 36, 391–409. doi: 10.1023/A:1020909529486

Bonawitz, E., Shafto, P., Gweon, H., Chang, I., Katz, S., and Schulz, L. (2009). "The double-edged sword of pedagogy: modeling the effect of pedagogical contexts on preschoolers' exploratory play," in *Proceedings of the Annual Meeting of the Cognitive Science Society, CogSci 2009* (Amsterdam: VU University Amsterdam).

Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., and Schulz, L. (2011). The double-edged sword of pedagogy: instruction limits spontaneous exploration and discovery. *Cognition* 120, 322–330. doi: 10.1016/j.cognition.2010.10.001

Boster, J. S., and Johnson, J. C. (1989). Form or function: a comparison of expert and novice judgments of similarity among fish. *Am. Anthropol.* 91, 866–889. doi: 10.1525/aa.1989.91.4.02a00040

Brown, C. H. (1984). *Language and Living Things: Uniformities in Folk Classification and Naming*. New Brunswick, NJ: Rutgers University Press.

Bruner, J. S., Goodnow, J. J., and Austin, G. A. (1967). *A Study of Thinking*. New York, NY: Science Editions.

Bunting, T. E., and Cousins, L. R. (1985). Environmental dispositions among school-age children: a preliminary investigation. *Environ. Behav.* 17, 725–768. doi: 10.1177/0013916585176004

Carey, S. (1985). *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Coley, J. D. (2012). Where the wild things are: informal experience and ecological reasoning. *Child Dev.* 83, 992–1006. doi: 10.1111/j.1467-8624.2012.01751.x

Coley, J. D., and Betz, N. (2018). "Expertise," in *The SAGE Encyclopedia of Lifespan Human Development*, eds M. H. Bornstean, M. E. Arteberry, K. L. Fingerman, and J. E. Lansford (Thousand Oaks, CA: Sage Publications, Inc).

Coley, J. D., and Muratore, T. M. (2012). "Trees, fish and other fictions," in *Evolution Challenges: Integrating Research and Practice in Teaching and Learning About Evolution*, eds K. S. Rosengren, S. K. Brem, E. Evans, and G. M. Sinatra (Oxford: Oxford University Press), 22–46. doi: 10.1093/acprof:oso/9780199730421.003.0002

Coley, J. D., Shafto, P., Stepanova, O., and Barraff, E. (2005). "Knowledge and category-based induction," in *Categorization Inside and Outside the Laboratory: Essays in Honor of Douglas L. Medin*, eds W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, and P. Wolff (Washington, DC: American Psychological Association), 69–85.

Coley, J. D., Solomon, G. E., and Shafto, P. (2002). "The development of folkbiology: a cognitive science perspective on children's understanding of the biological world," in *Children and Nature: Psychological, Sociocultural, and Evolutionary Investigations*, eds P. H. Kahn Jr. and S. R. Kellert (Cambridge, MA: MIT Press), 65–92.

Coley, J. D., and Vasilyeva, N. Y. (2010). "generating inductive inferences: premise relations and property effects," in *Psychology of Learning and Motivation*, Vol. 53, ed. B. Ross (Cambridge, MA: Academic Press), 183–226. doi: 10.1016/s0079-7421(10)53005-6

Coni, A. G., Ison, M., and Vivas, J. (2019). Conceptual flexibility in school children: switching between taxonomic and thematic relations. *Cogn. Dev.* 52:100827. doi: 10.1016/j.cogdev.2019.100827

Coulson, R. L., Feltovich, P. J., and Spiro, R. J. (1997). Cognitive flexibility in medicine: an application to the recognition and understanding of hypertension. *Adv. Health Sci. Educ.* 2, 141–161. doi: 10.1007/978-94-007-1793-0_8

Dane, E. (2010). Reconsidering the trade-off between expertise and flexibility: a cognitive entrenchment perspective. *Acad. Manag. Rev.* 35, 579–603. doi: 10.5465/amr.2010.53502832

Davidson, M. C., Amso, D., Anderson, L. C., and Diamond, A. (2006). Development of cognitive control and executive functions from 4–13 years: evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia* 44, 2037–2078. doi: 10.1016/j.neuropsychologia.2006.02.006

De Deyne, S., Kenett, Y. N., Anaki, D., Faust, M., and Navarro, D. J. (2016). "Large-scale network representations of semantics in the mental lexicon," in *Big Data in Cognitive Science: From Methods to Insights*, ed. M. N. Jones (Routledge: Taylor and Francis Group), 174–202.

Deak, G. O. (2003). The development of cognitive flexibility and language abilities. *Adv. Child Dev. Behav.* 31, 273–328. doi: 10.1016/S0065-2407(03)31007-9

Deák, G. O., and Wiseheart, M. (2015). Cognitive flexibility in young children: general or task-specific capacity? *J. Exp. Child Psychol.* 138, 31–53. doi: 10.1016/j.jecp.2015.04.003

Diamond, A. (2013). Executive functions. *Annu. Rev. Psychol.* 64, 135–168. doi: 10.1146/annurev-psych-113011-143750

Ferry, A. L., Hespos, S. J., and Waxman, S. R. (2010). Categorization in 3—and 4—month—old infants: an advantage of words over tones. *Child Dev.* 81, 472–479. doi: 10.1111/j.1467-8624.2009.01408.x

Fisher, A. V. (2015). Development of inductive generalization. *Child Dev. Perspect.* 9, 172–177. doi: 10.1111/cdep.12127

Garon, N., Bryson, S. E., and Smith, I. M. (2008). Executive function in preschoolers: a review using an integrative framework. *Psychol. Bull.* 134, 31–60. doi: 10.1037/0033-2909.134.1.31

Gelman, S. A. (2003). *The Essential Child: Origins of Essentialism in Everyday Thought*. New York, NY: Oxford University Press.

Gelman, S. A., and Coley, J. D. (1990). The importance of knowing a dodo is a bird: categories and inferences in 2-year-old children. *Dev. Psychol.* 26:796. doi: 10.1037/0012-1649.26.5.796

Gelman, S. A., and Coley, J. D. (1991). "Language and categorization: the acquisition of natural kind terms," in *Perspectives on Language and Thought: Interrelations in Development*, eds S. A. Gelman and J. P. Byrnes (Cambridge: Cambridge University Press), 146–196. doi: 10.1017/cbo9780511983689.006

Gelman, S. A., and Markman, E. M. (1986). Categories and induction in young children. *Cognition* 23, 183–209. doi: 10.1016/0010-0277(86)90034-x

Gelman, S. A., and Markman, E. M. (1987). Young children's inductions from natural kinds: the role of categories and appearances. *Child Dev.* 58, 1532–1541. doi: 10.2307/1130693

Glaser, B., and Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York, NY: Adline de Gruyter.

Graham, S. A., Gelman, S. A., and Clarke, J. (2016). Generics license 30-month-olds' inferences about the atypical properties of novel kinds. *Dev. Psychol.* 52:1353. doi: 10.1037/dev0000183

Heit, E., and Hayes, B. K. (2005). Relations among categorization, induction, recognition, and similarity: comment on Sloutsky and Fisher (2004). *J. Exp. Psychol. Gen.* 134, 596–605. doi: 10.1037/0096-3445.134.4.596

Hoehn, J. R., and Finkelstein, N. D. (2018). Students' flexible use of ontologies and the value of tentative reasoning: examples of conceptual understanding in three canonical topics of quantum mechanics. *Phys. Rev. Phys. Educ. Res.* 14:010122. doi: 10.1103/PhysRevPhysEducRes.14.010122

Inhelder, B., and Piaget, J. (1961/1999). *The Early Growth of Logic in the Child: Classification and Seriation*. Abingdon: Routledge.

Keates, J., and Graham, S. A. (2008). Category markers or attributes: why do labels guide infants' inductive inferences? *Psychol. Sci.* 19, 1287–1293. doi: 10.1111/j.1467-9280.2008.02237.x

Keil, F. C. ed. (1989). "The MIT Press series in learning, development, and conceptual change," in *Concepts, Kinds, and Cognitive Development* (Cambridge, MA: MIT Press).

Krieter, F. E., Julius, R. W., Tanner, K. D., Bush, S. D., and Scott, G. E. (2016). Thinking like a chemist: development of a chemistry card-sorting task to probe conceptual expertise. *J. Chem. Educ.* 93, 811–820. doi: 10.1021/acs.jchemed.5b00992

Lin, E. L., and Murphy, G. L. (2001). Thematic relations in adults' concepts. *J. Exp. Psychol. Gen.* 130:3. doi: 10.1037/0096-3445.130.1.3

López, A., Atran, S., Coley, J. D., Medin, D. L., and Smith, E. E. (1997). The tree of life: universal and cultural features of folkbiological taxonomies and inductions. *Cogn. Psychol.* 32, 251–295. doi: 10.1006/cogp.1997.0651

Markman, E. M., and Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: taxonomic versus thematic relations. *Cogn. Psychol.* 16, 1–27. doi: 10.1016/0010-0285(84)90002-1

Medin, D. L. (1989). Concepts and conceptual structure. *Am. Psychol.* 44:1469. doi: 10.1037/0003-066X.44.12.1469

Medin, D. L., Lynch, E. B., Coley, J. D., and Atran, S. (1997). Categorization and reasoning among tree experts: do all roads lead to Rome? *Cogn. Psychol.* 32, 49–96. doi: 10.1006/cogp.1997.0645

Medin, D. L., Ross, N. O., Atran, S., Cox, D., Coley, J., Proffitt, J. B., et al. (2006). Folkbiology of freshwater fish. *Cognition* 99, 237–273. doi: 10.1016/j.cognition.2003.12.005

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cogn. Psychol.* 41, 49–100. doi: 10.1006/cogp.1999.0734

Murphy, G. L., and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychol. Rev.* 92, 289. doi: 10.1037/0033-295x.92.3.289

National Research Council (2012). *A Framework for k-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academy Press.

Nguyen, S. P. (2007). Cross-classification and category representation in children's concepts. *Dev. Psychol.* 43:719. doi: 10.1037/0012-1649.43.3.719

Nguyen, S. P., and Murphy, G. L. (2003). An apple is more than just a fruit: cross—classification in children's concepts. *Child Dev.* 74, 1783–1806. doi: 10.1046/j.1467-8624.2003.00638.x

Noyes, A., and Keil, F.C. (2020). "There is no privileged link between kinds and essences early in development," in *Proceedings of the National Academy of Sciences* 117, 10633–10635.

Noyes, A., and Keil, F. C. (2019). Generics designate kinds but not always essences. *Proc. Natl. Acad. Sci. U.S.A.* 116, 20354–20359. doi: 10.1073/pnas.1900105116

Proffitt, J. B., Coley, J. D., and Medin, D. L. (2000). Expertise and category-based induction. *J. Exp. Psychol. Learn. Mem. Cogn.* 26, 811–828. doi: 10.1037/0278-7393.26.4.811

Rips, L. J. (1989). "Similarity, typicality, and categorization," in *Similarity and Analogical Reasoning*, eds S. Vosniadou and A. Ortony (Cambridge: Cambridge University Press), 21–59. doi: 10.1017/cbo9780511529863.004

Ritter, S. M., Damian, R. I., Simonton, D. K., van Baaren, R. B., Strick, M., Derks, J., et al. (2012). Diversifying experiences enhance cognitive flexibility. *J. Exp. Soc. Psychol.* 48, 961–964. doi: 10.1016/j.jesp.2012.02.009

Ross, B. H., and Murphy, G. L. (1999). Food for thought: cross-classification and category organization in a complex real-world domain. *Cogn. Psychol.* 38, 495–553. doi: 10.1006/cogp.1998.0712

Ross, N., Medin, D., Coley, J. D., and Atran, S. (2003). Cultural and experiential differences in the development of folkbiological induction. *Cogn. Dev.* 18, 25–47. doi: 10.1016/s0885-2014(02)00142-9

Scott, M. S., Serchuk, R., and Mundy, P. (1982). Taxonomic and complementary picture pairs: ability in two to five-year-olds. *Int. J. Behav. Dev.* 5, 243–256. doi: 10.1177/016502548200500206

Shafto, P., and Coley, J. D. (2003). Development of categorization and reasoning in the natural world: novices to experts, naive similarity to ecological knowledge. *J. Exp. Psychol. Learn. Mem. Cogn.* 29, 641–648. doi: 10.1037/0278-7393.29.4.641

Shafto, P., Coley, J. D., and Baldwin, D. (2007a). Effects of time pressure on context-sensitive property induction. *Psychon. Bull. Rev.* 14, 890–894. doi: 10.3758/bf03194117

Shafto, P., Coley, J. D., and Vitkin, A. (2007b). "Availability in category-based induction," in *Inductive Reasoning: Experimental, Developmental, and Computational Approaches*, eds A. Feeney and E. Heit (Cambridge: Cambridge University Press), 114–136. doi: 10.1017/cbo9780511619304.006

Sloutsky, V. M. (2010). From perceptual categories to concepts: what develops? Cognitive science, 34, 1244–1286. doi: 10.1111/j.1551-6709.2010.01129.x

Sloutsky, V. M., and Fisher, A. V. (2004). Induction and categorization in young children: a similarity-based model. *J. Exp. Psychol. Gen.* 133:166. doi: 10.1037/0096-3445.133.2.166

Sloutsky, V. M., Kloos, H., and Fisher, A. V. (2007). When looks are everything: appearance similarity versus kind information in early induction. *Psychol. Sci.* 18, 179–185. doi: 10.1111/j.1467-9280.2007.01869.x

Smiley, S. S., and Brown, A. L. (1979). Conceptual preference for thematic or taxonomic relations: a nonmonotonic age trend from preschool to old age. *J. Exp. Child Psychol.* 28, 249–257. doi: 10.1016/0022-0965(79)90087-0

Spiro, R. J. (1988). "Cognitive flexibility theory: advanced knowledge acquisition in ill-structured domains," in *Center for the Study of Reading Technical Report; no. 441* (Champaign, IL: University of Illinois at Urbana-Champaign).

Switzer, J., and Graham, S. A. (2017). 14- to 16-month-olds attend to distinct labels in an inductive reasoning task. *Front. Psychol.* 8:609. doi: 10.3389/fpsyg.2017.00609

Thibaut, J. P., Nguyen, S. P., and Murphy, G. L. (2016). Body and soul: do children distinguish between foods when generalizing biological and psychological properties? *Early Educ. Dev.* 27, 1250–1262. doi: 10.1080/10409289.2016.1146551

Unger, L., and Fisher, A. V. (2019). Rapid, experience-related changes in the organization of children's semantic knowledge. *J. Exp. Child Psychol.* 179, 1–22. doi: 10.1016/j.jecp.2018.10.007

Waxman, S. R., and Namy, L. L. (1997). Challenging the notion of a thematic preference in young children. *Dev. Psychol.* 33:555. doi: 10.1037/0012-1649.33.3.555

Williams, W. M., Papierno, P. B., Makel, M. C., and Ceci, S. J. (2004). Thinking like a scientist about real-world problems: the cornell institute for research on children science education program. *J. Appl. Dev. Psychol.* 25, 107–126. doi: 10.1016/j.appdev.2003.11.002

Zaitchik, D., Iqbal, Y., and Carey, S. (2014). The effect of executive function on biological reasoning in young children: an individual differences study. *Child Dev.* 85, 160–175. doi: 10.1111/cdev.12145

Zelazo, P. D. (2006). The dimensional change card sort (DCCS): a method of assessing executive function in children. *Nat. Protoc.* 1, 297–301. doi: 10.1038/nprot.2006.46

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Betz and Coley. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 29 October 2020
doi: 10.3389/fpsyg.2020.531640

Reasoning With Conditionals About Everyday and Mathematical Concepts in Primary School

Anastasia Datsogianni[1,2], Beate Sodian[2,3], Henry Markovits[4] and Stefan Ufer[1,2]*

[1] Chair of Mathematics Education, Department of Mathematics, LMU Munich, Munich, Germany, [2] REASON International Doctoral School, LMU Munich, Munich, Germany, [3] Chair of Developmental Psychology, Department of Psychology, LMU Munich, Munich, Germany, [4] Department of Psychology, Université du Québec à Montréal (UQAM), Montréal, QC, Canada

A research link between conditional reasoning and mathematics has been reported only for late adolescents and adults, despite claims about the pivotal importance of conditional reasoning, i.e., reasoning with if–then statements, in mathematics. Secondary students' problems with deductive reasoning in mathematics have been documented for a long time. However, evidence from developmental psychology shows that even elementary students possess some early conditional reasoning skills in familiar contexts. It is still an open question to what extent conditional reasoning with mathematical concepts differs from conditional reasoning in familiar everyday contexts. Based on Mental Model Theory (MMT) of conditional reasoning, we assume that (mathematical) content knowledge will influence the generation of models, when conditionals concern mathematical concepts. In a cross-sectional study, 102 students in Cyprus from grades 2, 4, and 6 solved four conditional reasoning tasks on each type of content (everyday and mathematical). All four logical forms, modus ponens (MP), modus tollens (MT), denial of the antecedent (DA), and affirmation of the consequent (AC), were included in each task. Consistent with previous findings, even second graders were able to make correct inferences on some logical forms. Controlling for Working Memory (WM), there were significant effects of grade and logical form, with stronger growth on MP and AC than on MT and DA. The main effect of context was not significant, but context interacted significantly with logical form and grade level. The pattern of results was not consistent with the predictions of MMT. Based on analyses of students' chosen responses, we propose an alternative mechanism explaining the specific pattern of results. The study indicates that deductive reasoning skills arise from a combination of knowledge of domain-general principles and domain-specific knowledge. It extends results concerning the gradual development of primary students' conditional reasoning with everyday concepts to reasoning with mathematical concepts adding to our understanding of the link between mathematics and conditional reasoning in primary school. The results inspire the development of educational interventions, while further implications and limitations of the study are discussed.

Keywords: conditional reasoning, primary school ages, domain knowledge, mathematics content, everyday content

OPEN ACCESS

Edited by:
Amanda C. Brandone, Lehigh University, United States

Reviewed by:
Patricia L. Mirabile, University of Amsterdam, Netherlands
Dan Kim, The Ohio State University, United States

****Correspondence:***
Anastasia Datsogianni
anastasia.datsogianni@psy.lmu.de;
datsogianni.a@unic.ac.cy

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *31 January 2020*
Accepted: *05 October 2020*
Published: *29 October 2020*

Citation:
Datsogianni A, Sodian B, Markovits H and Ufer S (2020) Reasoning With Conditionals About Everyday and Mathematical Concepts in Primary School. Front. Psychol. 11:531640. doi: 10.3389/fpsyg.2020.531640

INTRODUCTION

The ability to make valid deductions is considered of central importance for scientific reasoning, hypotheses generation, and evaluation (Kuhn et al., 1988), as well as for mathematical thinking (Moshman, 1990; Markovits and Lortie-Forgues, 2011) and learning and success (Nunes et al., 2007; Morsanyi and Szücs, 2014). Crombie (1994) puts forward mathematical deduction as one of six styles of scientific reasoning, which Kind and Osborne (2017) propose as a framework for science education. Indeed, deductive reasoning not only plays an important role for reasoning in mathematics, but it is also a core method to derive scientific hypotheses as conclusions from central assumptions within scientific theories or to draw conclusions from an experiment based on general scientific principles. Kind and Osborne (2017) argue that an exclusive focus of psychological and science education research on a restricted set of scientific reasoning styles, such as experimentation, offers students an "impoverished account of scientific thinking." In this contribution, we focus on deductive reasoning with mathematical concepts as an important mode of scientific reasoning, which extends and complements research on other scientific styles, such as experimentation.

The term *deductive reasoning* refers to forms of reasoning that ensure that, if the premises are true, the conclusion is necessarily true as well. In particular, in mathematics, deduction is strongly associated with the development of students' ability to understand formal proofs (Foltz et al., 1995), and it is considered as the central mode of reasoning in mathematical theories. Even basic mathematical concepts are characterized by a strict logical structure, which makes deduction a central mode of reasoning in mathematics at every educational level. One of the key components of deductive reasoning is conditional reasoning, i.e., reasoning with "if–then statements." Mathematical concepts are characterized by specific properties, which often have the form of conditional statements (e.g., "If the sum of two whole numbers is odd, then their product is even"). Making inferences with such statements requires conditional reasoning skills. Conditional statements are not only a central part of mathematical discourse, but they also occur frequently in everyday language and communication, for example, as with rules such as "If you have a fever, you will have to stay in bed."

When it comes to conditional reasoning with mathematical concepts, we can assume that knowledge of these concepts develops substantially during primary school. On the contrary, the everyday conditional rules usually applied in research with primary school children (e.g., the fever example above) can be considered to be familiar to them. Of course, detailed knowledge about the causal mechanisms behind these rules reflects substantial knowledge about scientific concepts (e.g., biology in the example above), but the phenomena connected to the rules (having fever, staying in bed, and recovering) are most likely well known to primary school children. For mathematical content (MA), however, we see a parallel development of knowledge about mathematical concepts and connected phenomena, as well as a general improvement in conditional reasoning skills at this age level. It is an open question, however, to what extent conditional reasoning with these contents also requires mathematical knowledge about the concepts involved in the conditional statement. This taps into the question to what extent scientific (and, in particular, conditional) reasoning is guided by general abstract rules that can be applied without much knowledge about the concrete content of the topic at hand and to what extent it relies on explicit knowledge of this content. Conditional reasoning in mathematics has been investigated primarily with adults (Durand-Guerrier, 2003; Stylianides et al., 2004; Inglis and Simpson, 2009; Attridge and Inglis, 2013) and secondary school students, with results suggesting difficulties in reasoning with conditionals in mathematics (e.g., Küchemann and Hoyles, 2002). Despite the potential importance of this form of reasoning, there is very little evidence about elementary students' abilities to reason with conditionals with MAs. Although previous results have shown that secondary school students have real difficulties with mathematical conditional reasoning (e.g., Küchemann and Hoyles, 2002), psychological studies show that even very young children can reason correctly with conditionals with familiar everyday contents (EDs) (e.g., Markovits and Thompson, 2008). This leads to the question if and how conditional reasoning with mathematical concepts is different from conditional reasoning with EDs.

The present study focuses on primary school students from grades 2, 4, and 6, an age range in which basic mathematical concepts such as arithmetic operations are acquired, which makes it possible to study conditional inferences about these concepts. We will specifically compare elementary students' conditional reasoning skills about everyday situations and about mathematical concepts. At this age, conditional reasoning skills with everyday contexts have been found in the past (e.g., Markovits and Thompson, 2008; Markovits, 2017), but it is an open question to what extent such skills can be transferred to reasoning about mathematical concepts, which are acquired at this age.

Conditional Inferences

Conditional reasoning refers to making inferences based on a conditional statement of the form "if p, then q," which is called the *major premise* in a conditional reasoning task. In this setting, p is called the *antecedent*, and q is called the *consequent*. Conditional inferences require a further, *minor premise*. Four different minor premises differentiate four possible logical forms. These four forms can be described systematically by the wording of the minor premise (positive vs. negative) and the type of normative correct conclusion (**Table 1**). For example, when the major premise is "If the sum of two whole numbers is odd, then their product is even," a minor premise could be "we have two numbers, which do not have an odd product" (i.e., they have an even product). This premise is negatively worded and allows a definite conclusion: "The sum of the two numbers is not even"). Based on the traditional interpretation of conditionals that p is sufficient, but not necessary for q (Evans and Over, 2004), and depending on the logical form (minor premise), different conclusions can be drawn. Indeed, one can prove that the sum of two whole numbers being odd is sufficient for their product to be even. However, an odd sum is not necessary to arrive at

TABLE 1 | Logical forms in conditional reasoning for the major premise "if *p*, then *q*."

Name of form (abbreviation)	Minor premise	Normatively correct conclusion	Minor premise wording	Conclusion type	Minimally required models to make a correct inference
Modus ponens (MP)	"*p* is true"	"so *q* is true"	Positive	Definite	"*p* and *q*"
Modus tollens (MT)	"*q* is false"	"so *p* is false"	Negative	Definite	"*p* and *q*," "not-*p* and not-*q*"
Affirmation of the consequent (AC)	"*q* is true"	"so *p* or not *p*"	Positive	Indefinite	"*p* and *q*," "not-*p* and *q*"
Denial of the antecedent (DA)	"*p* is false"	"so *q* or not *q*"	Negative	Indefinite	"not-*p* and not-*q*," "not-*p* and *q*"

See ***Supplementary Tables 1, 2*** *in the* ***Supplementary Materials (A.1)*** *for explicit examples using conditionals with everyday and MA.*

an even product, as, for example, in $4 \times 6 = 24$ (even product), but $4 + 6 = 10$ (even sum). Definite conclusions are possible for modus ponens (MP; minor premise "*p* is true" or "we have two numbers that have an odd sum" in the example) and modus tollens (MT; minor premise "q is false" or "we have two numbers that do not have an even product"). Acceptance of consequent (AC; minor premise "q is true" or "we have two numbers that have an even product") does not allow definite conclusions about *p* and *q*. In this case, the conditional does exclude that neither the sum of the two numbers is odd (e.g., 3 and 4), nor it is even (e.g., 2 and 4). Thus, the correct conclusion is indefinite: We cannot say whether the antecedent is true or not. In the same way, denial of antecedent (DA; minor premise "*p* is false" or "we have two numbers that do not have an odd sum") does not allow a definite conclusion about *q* (or the product of the two numbers).

MMT of Conditional Inference

One of the most influential theories that have been used to describe conditional reasoning in young children is Mental Model Theory (MMT). This suggests that conclusions are drawn by constructing mental models that encode information about the meaning of the conditional (Johnson-Laird and Byrne, 2002). Such models are generated from a semantic analysis of rules and represent possible states of affairs under these rules. If a given model represents a potential counterexample to a putative conclusion, this conclusion will be denied; otherwise, it will be accepted. Within this perspective, conditional reasoning depends on knowledge about the specific content of a conditional. In other words, individuals use the meaning of premises and their knowledge about the content to think about what is possible given the premises (Nickerson, 2015). Prior studies have shown that MMT accurately describes conditional reasoning processes among elementary students using everyday content (e.g., Markovits, 2000; Markovits and Thompson, 2008). Thus, we will use MMT in the following as the basis for our analyses of conditional reasoning in the primary school age.

According to MMT, humans generate "*p* and *q*" as a standard model for every conditional. To make a valid deduction on MP tasks, an instance of this base model is sufficient: the minor premise *p* is part of the model, and the other part of the model *q* can be used as the correct conclusion. To arrive at a valid conclusion for MT tasks, an additional model "not-*p* and not-*q*" is necessary, because the standard model "*p* and *q*" is not compatible with the minor premise "not-*q*." To derive that AC does not allow a valid conclusion, minimally models "*p* and *q*" and "not-*p* and *q*" are necessary: Both models are compatible with the minor premise "not-*p*," but offer different conclusions "*q*" and "not-*q*." Similarly, the models "not-*p* and not-*q*" and "not-*p* and q" are required for DA inferences. Critical to correctly making indefinite conclusions on the AC and DA forms is the ability to generate models of the form "not-*p* and *q*." These refer to alternative antecedents, which are counterexamples for the typical errors in AC and DA inferences. Another class of alternative models are instances of "*p* and not-*q*," which are counterexamples for the correct MP and MT inferences, called disablers (Cummins et al., 1991). Disablers are not compatible with the conditional itself and thus may lead to wrong conclusions. Higher availability of disablers is related to a lower rate of MP and MT acceptances (Cummins, 1995; Janveau-Brennan and Markovits, 1999), whereas higher availability of alternative antecedents is related to higher rates of correct reasoning in AC and DA forms (Markovits and Vachon, 1990; Cummins et al., 1991). Janveau-Brennan and Markovits (1999) found that conditional reasoning in young children (ages 7–12 years) is affected by rates of both alternative and disabler generation in a way that is similar to their effect in adults. In addition, studies indicate that correct DA and AC reasoning correlates negatively with correct MP and MT reasoning (Newstead et al., 2004; Morsanyi et al., 2018). This could reflect a connection between alternative generation (supporting correct DA and AC reasoning) and disabler generation (leading to incorrect MP and MT reasoning).

Development of Conditional Reasoning

Based on MMT, we can expect the development of conditional reasoning to depend on at least two mechanisms, which will be laid out in the next paragraphs: (1) the acquisition of general schemata to describe conditional statements, which guide adequate interpretation of the different logical forms and provide strategies of model generation independently of the specific content of the conditional, and (2) an increase in knowledge about the conditionals' contents, which is necessary to build up mental models in general and more specifically to construct disablers and alternatives. While the first mechanism can be assumed to have effects independent of the content of the conditionals, the second mechanism allows for the construction or retrieval of mental models for specific conditionals, which is content dependent.

Content-Independent Mechanisms

Within MMT, the general ability to construct interpretations of conditionals that are more complex has been hypothesized to

underlie development. Specifically, more complex interpretations require maintaining additional models in memory, which requires increased working memory capacity. According to Barrouillet and Lecas (1999), the development of such general schemata of conditionals starts from the conjunctive-like interpretation (model: "*p* and *q*"), developing to the biconditional interpretation (models: "*p* and *q*"; "not-*p* and not-*q*") and then to the full conditional interpretation (complete three-model representation: "*p* and *q*"; "not-*p* and not-*q*"; "not-*p* and *q*"). They suggest that the development of conditional reasoning abilities is determined by a developmental increase in working memory capacity. Whereas this model suggests relatively sharp developmental differences, other results show a more gradual change. For example, Janveau-Brennan and Markovits (1999) found a steady age-related development in the ability to make correct AC and DA inferences between grades 1 and 6, as well as a gradual increase in retrieval of disabling conditions leading to less correct MP reasoning. The latter is explained by an erroneous application of disablers to MP inferences. Many studies have found that the AC and DA forms are usually not mastered before the age of 11–12 years, whereas even only about one-third of adults have been found to systematically make these inferences normatively (Gauffroy and Barrouillet, 2009; Ricco, 2010; Moshman, 2011; Markovits, 2014; Christoforides et al., 2016). In addition, many studies (e.g., Janveau-Brennan and Markovits, 1999) have found that specific content strongly affects conditional reasoning. Summarizing, the current evidence indicates that, possibly connected with working memory capacity, children acquire schemata of conditional reasoning, which allow correct MP reasoning first, then MT, and later AC and DA reasoning. However, as we shall see, there are clear indications that, because of the necessity to retrieve or generate alternatives and disablers, these reasoning skills are subject to important content effects.

Regarding the development of conditional reasoning with positively versus negatively worded minor premises (i.e., MP and AC vs. MT and DA), the literature provides less information. The negations involved in negatively worded minor premises have been hypothesized to pose specific difficulties (Schroyens et al., 2001). MMTs usually assume that mental models only represent *possibilities* that can occur given the premises—not what is impossible given the premises (principle of truth, Johnson-Laird, 2001), which could lead to problems if a negation leads to an unspecified situation (e.g., while "not wet" has a similar meaning as "dry," "blue" has no specific opposite standing for "not blue"). The latter case would require an abstract understanding of negations and thus can be seen as a content-independent part of conditional reasoning skills. However, the account for the effects of positive versus negative wording in the literature is not as explicit as for definite versus indefinite forms.

Content-Specific Mechanisms and Knowledge About Conditionals' Contents

Apart from general individual development, the specific content of the conditional has been found to influence the ability to make conditional inferences. Previous studies (Markovits, 2000; Markovits and Thompson, 2008) have shown that even 6- or 7-year-old children can reason logically with the AC and DA inferences, when the content refers to simple categorical premises (e.g., "If an animal is a cat, then it has legs"). In particular, in the study by Markovits and Thompson (2008), 6-, 7-, and 9-year-old students were observed to make valid conclusions about MP and AC inferences in a familiar categorical context (e.g., "If something is a car, then it has a motor. Now, suppose that you see a car" under either categorical instructions ("Is it certain that it has a motor?") or probabilistic instructions ("How sure is it that it has a motor?"). Chao and Cheng (2000) also found evidence of the domain specificity of conditional reasoning in preschool children. In detail, the latter study examined preschoolers' conditional reasoning skills within permission and arbitrary concepts showing that pragmatic (permission) conditional rules (e.g., "If it is windy, then he must not have shorts on") seem to emerge earlier than formal (arbitrary) conditional rules (e.g., "If it is windy, then there must not be an orange in the box") as MP and MT inferences were approached reliably by the students only in the permission context. These kinds of effects have led to the conclusion that conditional reasoning might be domain-specific (e.g., Cummins, 1996b; Chao and Cheng, 2000), especially in early ages. In fact, based on a series of studies by Markovits and Lortie-Forgues (2011) and Markovits (2014), which examined the context effects as well as the alternative generation effects on conditional reasoning skills among individuals from age 7 to age 19 years, a clear developmental pattern was proposed. This pattern (Markovits, 2014) suggests that 7- and 8-year-old students possess conditional reasoning skills with categorical premises, 10- to 12-year-old children can make logical conditional inferences with familiar causal premises; 14–16 year olds can do so with causal and counterfactual premises, whereas adults older than 20 years also perform well with abstract conditionals.

That conditional reasoning performance depends on the content of the conditionals is in line with core assumptions of MMT. According to this, mental models need to be retrieved or constructed based on knowledge about the situation contained in the conditional statement. As previously mentioned, studies on the effects of content on conditional reasoning have concentrated on broad categories, which have been shown to affect in particular retrieval of alternative antecedents (e.g., Markovits and Lortie-Forgues, 2011). However, the effects of more specific forms of content variation, such as that involved in reasoning with mathematical concepts, have not often been studied.

Conditional Reasoning About Mathematical Concepts

One possible type of content, for which knowledge is acquired during primary school age, comprises mathematical concepts. Consider, for example, the conditional "If a house has three floors with four windows each, then it has 12 windows." Representing this situation *per se* does not require substantial mathematical knowledge beyond representing cardinal numbers, which is usually acquired by early primary school age (e.g., Litkowski et al., 2020). For example, representing a model of the type "*p* and *q*" would consist of an instance of a house that has three floors with four windows each. Calculating that there are 12 windows overall is a basic procedure, which can be solve by primary school students with a variety of basic strategies (e.g., addition $4 + 4 + 4$,

or fact retrieval of $3 \times 4 = 12$). However, generating an explicit alternative would involve imagining a different configuration of floors and windows per floor, which does consist of 12 windows (e.g., two floors with six windows each). This involves solving a reverse task, specifically to find a different configuration (e.g., two floors with six windows each) that also leads to 12 windows overall. Finding pairs of factors that have a given product requires substantial mathematical knowledge about multiplication, which has been found to develop slowly even until the end of primary school (e.g., Robinson et al., 2018).

Only sparse evidence about conditional reasoning with mathematical concepts is available for primary school. The mathematics education literature has focused mostly on older learners and has shown, for example, that dealing with mathematics and participating in mathematics instruction can lead to improved conditional reasoning skills (Handley et al., 2004; Inglis and Simpson, 2008, 2009; Attridge and Inglis, 2013; Morsanyi et al., 2017; Morsanyi et al., 2018). For secondary school students, research has focused mainly on conceptual issues such as the differentiation between a statement and its converse and less on drawing conditional inferences (Küchemann and Hoyles, 2002).

Given that both general schemata for conditional reasoning as well as mathematical knowledge develop during primary school age, it is of substantial interest to understand how these two developments interact. It can be assumed that knowledge about the content of familiar conditionals is widely available at this age, whereas knowledge about mathematical concepts varies substantially (Robinson et al., 2018). This would suggest that reasoning with mathematical concepts would be more difficult than reasoning with everyday statements. This would be consistent with prior results showing early reasoning skills with everyday content (e.g., Markovits and Thompson, 2008) and reports on secondary school students' problems dealing with mathematical conditionals (Küchemann and Hoyles, 2002). Different mechanisms could be hypothesized to explain such differences. First, the mathematical content of the conditional might affect decoding and representation of the conditional, which would be reflected in a relatively coherent performance difference between everyday reasoning and reasoning with mathematical content over all logical forms. This effect could also be moderated by the development of mathematical knowledge during primary school age and thus be larger for younger students. Beyond the initial problem representation, the retrieval or construction of alternative mental models is a second point in the MMT account of conditional reasoning that is particularly dependent on content-related knowledge. MMT would predict stronger content-related differences here for the two indefinite forms AC and DA, because more models are required to make valid inferences on these forms than for the definite forms MP and MT. Indeed, prior research has found an influence of alternative generation skills on AC reasoning rather than on MP reasoning (Janveau-Brennan and Markovits, 1999; Markovits and Quinn, 2002). Finally, because this effect depends on available knowledge about the conditional's content, it should be more pronounced for students in lower grades, leading to an interaction of content, logical form, and grade level. Given the lack of available evidence on elementary school students' conditional reasoning skills with mathematical concepts, more substantial hypotheses are hard to derive. However, the role of alternative generation found in prior studies with elementary school students (e.g., Markovits, 2017), together with progress in mathematical knowledge during elementary school age (Robinson et al., 2018), speaks to expecting the interaction of content, logical form, and grade level.

The Current Study

In this study, we contrast conditional reasoning about premises involving mathematical concepts with reasoning about familiar causal premises. To this end, we study conditionals about easily accessible situations that contain mathematical structures (e.g., numbers of windows in configurations for floors and windows per floor as in the example above). We chose structures related to multiplication and addition, concepts that are introduced in the first years of primary school. We assume that increasing knowledge about these concepts will affect conditional reasoning performance on top of the well-described development of conditional reasoning skills with familiar premises.

The main goal of this study is to investigate to what extent reasoning about mathematical concepts specifically affects primary school students' conditional reasoning in the four different logical forms and its development. Beyond replicating findings on the development of logical reasoning with familiar everyday statements, the following questions are addressed:

(1a) Is there a general disadvantage of reasoning with mathematical content (MA), compared to everyday content (ED), for primary age students?

Understanding the situations, in which we embedded the mathematical content for the conditionals, did not require substantial mathematical knowledge. Thus, we did not put forward a specific hypothesis about whether this factor would show a main effect.

(1b) Is such a general disadvantage larger for children in lower grades, as compared to upper primary school grades?

Again, because we embedded the mathematical concepts and structures in easily accessible situations, we also did not put forward a specific hypothesis about this specific interaction.

(2a) Is there a specific disadvantage of reasoning with mathematical concepts, as compared to everyday conditionals, for the indefinite logical forms AC and DA?

The role of alternative generation for reasoning with indefinite forms has been shown for everyday reasoning. Moreover, generating alternatives for the mathematical structures reflected in our conditionals requires well-connected mathematical knowledge. Thus, we expected a corresponding interaction of content and the type of inference (definite for MP and MT vs. indefinite for AC and DA).

(2b) Is there a disadvantage of a negative (vs. positive) wording on conditional reasoning on both contexts, for definite and indefinite conditionals?

Given the sparse evidence on the effects of positive versus negative wording, we approached this question in an exploratory manner, expecting lower performance for negative wordings because of the specific difficulties involved in dealing with negations.

(2c) Are the effects under (2a) and (2b) dependent on students' grade level?

Because knowledge about additive and multiplicative structures develops over primary school age, we expected that the effect discussed under (2a) would be more pronounced for students in lower grades, as compared upper primary school grades.

MATERIALS AND METHODS

Participants

Around 300 students and their parents were approached for participation in this cross-sectional study. A total of 102 elementary students (average age, 10 years 1 month) from grades 2, 4, and 6 living in Cyprus participated. Regarding the participants' socioeconomic background, the median category on the "books at home" question (Paulus, 2009) was "one complete bookcase (26–100 books)" in all grades, and distribution over the five answer alternatives (from "no or very few books" to "over 200 books") did not differ significantly over grades [$\chi^2(8) = 13.4$, $p = 0.10$]. Parents' written consent and children's oral assent were obtained for all participants. Further information about the sample is displayed in **Table 2**.

Design

Each participant took part in one individual 45- to 60-min face-to-face interview during regular school hours with the first author in a separate room of the school. The factors relating to *content* (everyday vs. mathematical) and logical form (*positive vs. negative* wording of minor premise, *definite* vs. *indefinite* conclusion, and the interaction of the two) were varied within subject, with randomized sequence of two content blocks, randomized sequence of conditionals within each content block, and randomized sequence of four minor premises (each relating to a different logical form) within each conditional.

Procedure

Initially, students were asked for their age, native language, and approximate number of books at their house ("books at home" question; Paulus, 2009). Participants were clearly informed about the anonymity and confidentiality of their replies, as well as their voluntary participation, clarifying that they were free to withdraw from the interview process anytime without any negative consequences. Then, participants were familiarized with the three answer alternatives ("yes, this is certainly so," "no, this is certainly not so," "you cannot say for sure, whether it is so, or not") and shown how to select their answers on the tablet computer screen. The answer options were represented by symbols on the tablet screen: A green check mark for "yes, this is certainly so," a red "x" for "no, this is certainly not so," and a question mark for "you cannot say for sure, whether it is so, or not". A short game with three questions about hidden marbles was used to check for comprehension of these answer alternatives (see **Supplementary Materials A.3**). An explanation was given in case of wrong answers. Afterward, two blocks (everyday vs. mathematical) of four conditionals each were presented (10–15 min per block). Each conditional was presented separately, and students were asked to make four conclusions based on four different minor premises, corresponding to the four logical forms. Then, a block of alternative generation tasks followed, which is not examined in this article. In the end of the interview, students' responses to a short working memory test and an arithmetic calculation test (not examined in this paper) were gathered.

All tasks were displayed using a tablet-based interview system, and children were expected to select their preferred answer by touching the screen on the respective part of the visual representation of answer alternatives. The interview system also randomized the sequence of blocks, conditionals, and minor premises. By ensuring this full randomization of the questions' order, we systematically controlled for possible order effects. No justifications for the answers were requested as we were interested in students' intuitive responses.

Conditional reasoning tasks

Eight conditionals were used to measure conditional reasoning skills (four conditionals per content condition). The verbal structure of the tasks was parallel in both content blocks. All conditionals were presented verbally and in a written form on a tablet computer. Participants were told for each conditional that they should assume that it was really true. For each of the four conclusions to be made on each conditional, students were presented with the major and the minor premise verbally and on the screen. They were asked if they could conclude that a given conclusion was *true for sure*, if it was *not true for sure*, or if *no definite conclusion* was possible. For example, for the conditional "If someone's finger is cut deeply while cooking, then it bleeds," the minor premise "George's finger is not bleeding" would have been presented to test the logical form MT. The students would have been asked, "Based on what he knows, what can Peter [the central character in our cover story] say for sure?" and the answer options were "George's finger has just been cut deeply while cooking" (yes), "George's finger has not just been cut deeply while cooking" (no), and "He cannot be sure whether George's finger has just been cut deeply while cooking or not" (uncertain).

The everyday conditional reasoning tasks contained familiar causal conditionals (with the antecedent and the consequent

TABLE 2 | Sample size, mean, and standard deviation (SD) of age in years and working memory scores by grade level.

Grade	*n*	Age in years, mean (*SD*)	Working memory score, mean (*SD*)
2	32	8.22 (0.72)	2.16 (0.99)
4	33	9.94 (0.53)	3.18 (1.04)
6	37	11.82 (0.53)	4.19 (1.66)

being the cause and effect, respectively: If a glass is dropped in the kitchen, then there is a sound; if someone's finger is cut deeply while cooking, then it will bleed; if someone jumps into a pool, then they will get wet; if someone breaks their arm, then they will hurt). The verbal structure of the introduction, conditionals, and answer alternatives were based on previous studies on conditional reasoning (e.g., Markovits and Lortie-Forgues, 2011).

The conditionals with mathematical content dealt with situations that contained mathematical structures. The specific structures and related concepts were multiplication and addition, because these concepts are included in the national curriculum up to grade 2. Comprehension questions were included for the conditionals, to control if students understood the situation in which the major rule was embedded.

For example, one of these situations was introduced in the following way:

"Peter is walking with the little explorers, and they just found some treasure boxes. We know that the boxes contain some blue and red diamonds. Each blue diamond is worth three gold coins. Each red diamond is worth two gold coins."

The corresponding comprehension question was: "In a treasure box, there is one blue diamond and two red diamonds. How many gold coins is this worth?" In case the child provided the correct answer (7), the reasoning tasks followed. After a wrong answer, the researcher repeated the explanations and posed the comprehension question for a second and last time. The answer was recorded, and the researcher continued without providing any feedback or hints to students. For mathematical conditionals, only answers on reasoning tasks for which the corresponding comprehension question was answered correctly were included in the analyses. In grade 2, answers to 62 of 128 presented conditionals (48.4%) were excluded, with eight students being excluded on all conditionals. In grade 4, 26 of 132 conditionals (19.7%) were excluded, and in grade 6, 15 of 148 conditionals (10.1%). In grades 4 and 6, at least two conditionals were included for each student.

The conditional (major rule) in the example before was "If the box contains exactly two blue diamonds and three red diamonds, then the diamonds in the box are worth 12 gold coins. It is certain that this is really true." For example, the logical form MT was presented in the following way: "This is Stelios. The diamonds in his box are not worth 12 gold coins. Based on what he knows, what can Peter say for sure?" The alternative answers were parallel to the everyday conditionals: "The box contains exactly two blue diamonds and three red diamonds" (yes), "The box does not contain exactly two blue diamonds and three red diamonds" (no), and "He cannot be sure whether the box contains exactly two blue diamonds and three red diamonds or not" (uncertain).

Overall, the students worked on 32 conditional reasoning tasks (2 content types × 4 conditionals each × 4 logical forms each). All tasks were tested through a pilot study ensuring their appropriateness for this age range (Datsogianni et al., 2018).

Working memory test

Working memory capacity was chosen as a control variable, because it has been found to predict mathematics skills (Holmes and Adams, 2006), as well as logical reasoning skills (Nakamichi, 2007, 2011). To measure working memory capacity, a backward digit span test (Wechsler Intelligence Scale for Children IV Digit Span Subtest) was used, as in previous studies on young children's conditional reasoning skills (e.g., Nakamichi, 2007, 2011). Specifically, sequences of digits were read out loud, and students were asked to repeat them in the reversed order. In the beginning they were provided with an example of a sequence of three digits (e.g., 9–2–7) and were asked to reproduce it backward; in case of a correct reply, the test continued. In case of a wrong reply, the correct response was given (7–2–9), and then a second example was presented. Regardless of the reply to the second example, no feedback on this was given. The test consisted of 14 trials. The first two trials contained two digits, each (2–5, and 6–3). The number of digits increased by one after every second trial. The test was discontinued after failure on two trials with the same number of digits. No hints were given on any of the trials. Each trial was scored 2 (if the child passed both trials), 1 (if the child passed only one item), or 0 (if the child failed both trials).

Analyses

The answers to the conditional reasoning tasks were analyzed using generalized linear mixed models, a generalization of logistic regression. It allows analyzing the data on item level, but still takes into account dependencies between answers provided by each student and on each task. The package lme4 in R was used (Bates et al., 2015). Grade level (grades 2, 4, and 6) was included as a between-subject factor, and content (everyday vs. mathematical), wording of the minor premise (positive vs. negative), and type of conclusion (definite vs. indefinite) as within-subject factors. Wald χ^2 tests were used to compare models during model selection and to analyze omnibus effects. Planned contrasts of estimated marginal means were used to compare performance in different cells of the experimental design. These contrasts are expressed in logarithmic odds ratios. For example, a main effect contrast of $b = 0.50$ can correspond to differences in solution rates of up to

$$\frac{e^{0.77+\frac{0.50}{2}}}{1+e^{0.77+\frac{0.50}{2}}} - \frac{e^{0.77-\frac{0.50}{2}}}{1+e^{0.77-\frac{0.50}{2}}} = 10.8\%$$

around the mean solution rate of the items in our study (a logit of 0.77 corresponds to the mean solution rate of 59.5% in our study). However, corresponding differences in solution rates can be substantially smaller, when very easy or very difficult task variants are compared. Bonferroni correction was applied when analyzing multiple contrasts along the same factors.

Ethics statement

The ethics approval was obtained from the Centre of Educational Research and Evaluation of Cyprus Pedagogical Institute, as well as the Cyprus Ministry of Education and Culture. Back-and-forth professional translation, from the original English language of the interview protocol into the Greek language and back, was conducted. Parents and students were informed that the participation in the study was completely voluntary, that answers would be handled confidentially, and that they could stop their participation at any time without any further consequences.

RESULTS

Descriptive Results

Overall, 59.5% of students' answers in the conditional reasoning tasks were correct. **Figure 1** shows the frequency of each answer option by conditionals' content and logical form. MP tasks were mostly answered correctly with affirmative conclusions ("q is true"), and correct disaffirmative conclusions ("p is false") were observed in more than 50% of the answers to MT tasks. Wrong indefinite conclusions ("uncertain") occurred rarely for MP tasks, but more often for MT tasks. For AC tasks, the correct indefinite conclusions ("one cannot say whether p is true or not") were about as frequent as the wrong affirmative conclusions ("p is true"). For DA tasks, correct indefinite conclusions ("one cannot say whether q is true or not") occurred less often than wrong disaffirmative conclusions in the everyday context, while the frequencies were comparable for the mathematical context. Summarizing, performance was highest in MP tasks (correct answer affirmative), followed by MT tasks (correct answer disaffirmative), whereas AC and DA tasks showed descriptively similar frequencies of correct indefinite conclusions ("uncertain"), with a slightly lower performance on everyday DA tasks.

For all combinations of grade levels, logical forms, and contents, the distribution of responses over the three possible options (affirmative, disaffirmative, and indefinite) differed significantly from a uniform distribution in all conditions, except for grade 2 on items with MA and the two negative forms (Bonferroni correction for 24 tests, $p = 0.02$ for MT and $p = 0.10$ for DA). Thus, that students might have applied a systematic guessing strategy must be taken into account for these two conditions.

Model Selection

In an initial step of model selection, we decided on the random intercepts and slopes to be included. Given the low number of conditionals per person, we analyzed only the random slopes for main effects of grade level and logical form over the different conditionals. Initial analyses indicated that including random slopes for interaction effects over individuals lead to singular

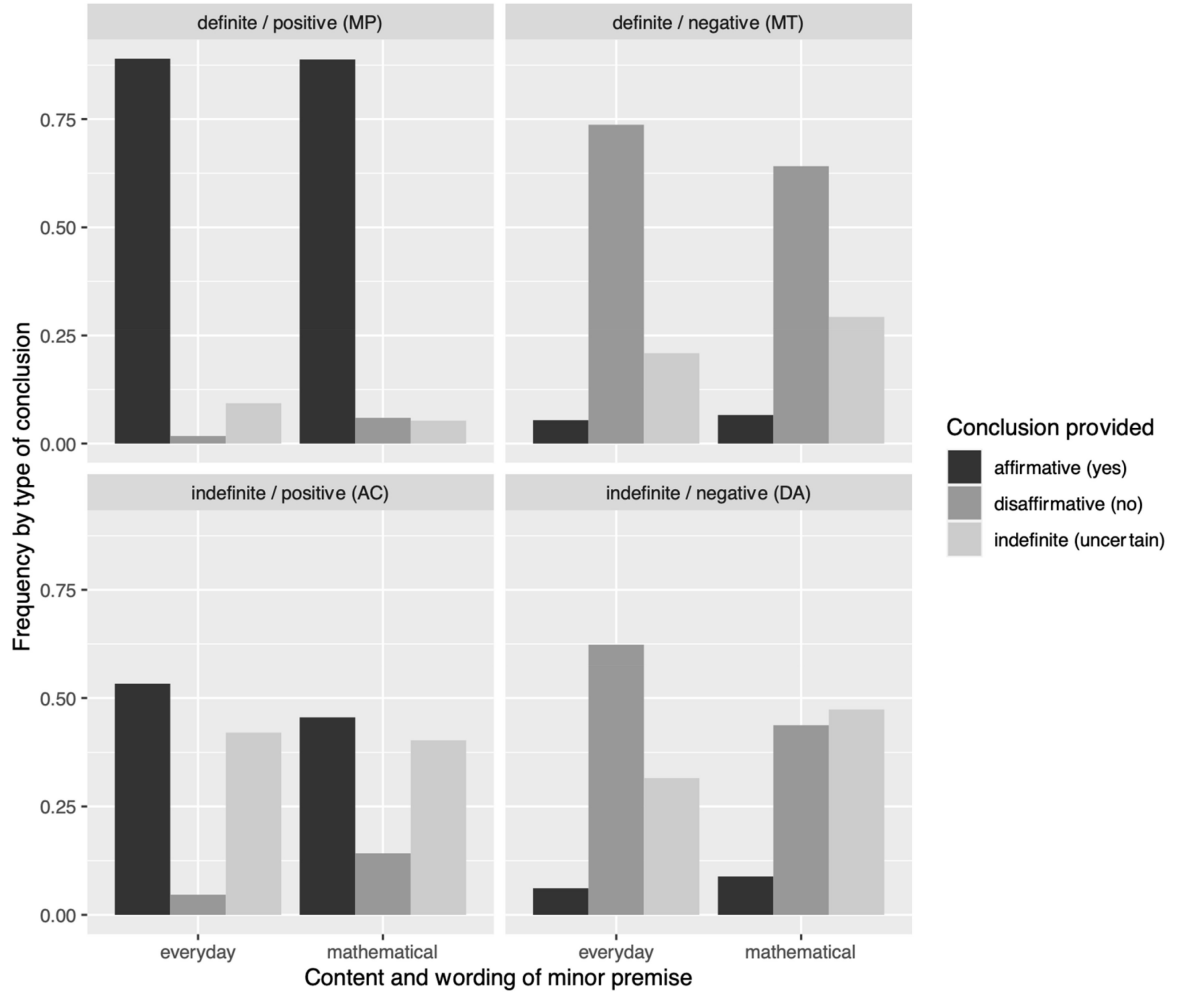


FIGURE 1 | Frequency of answers given by conditional content and logical form. ED, everyday content; MA, mathematical content.

model fit. Because singular models are prone to misinterpretation (Bates et al., 2018), we decided to analyze only random slopes for main effects of content and logical form over individuals. χ^2 difference tests indicated that leaving out the random slope for grade level over conditionals from a model containing random intercepts and random slopes for all main effects over conditionals and persons did not affect model fit significantly. However, removing the random slopes for the remaining main effects over persons and removing the random slopes for the remaining main effects over conditionals each affected model fit significantly. Thus, we decided to select the model with random intercepts and random slopes for *wording of the minor premise* (positive vs. negative) and *type of conclusion* (definite vs. indefinite) over conditionals, as well as random intercepts and random slopes for *wording of minor premise*, *type of conclusion*, and *content* over individuals. In a last step, we removed all non-significant interactions of fixed effects, which were not part of other significant interactions. This did not reduce model fit significantly, as well. Results of the χ^2 difference tests and the lme4 formulas for the models considered during model selection are given in the **Supplementary Materials A.2**.

χ^2 statistics for the fixed main and interaction effects in the final model are given in **Table 3**. Working memory as a control variable did not predict conditional reasoning scores significantly (**Table 3**). Estimated marginal means and confidence intervals are available in **Figure 2**.

General Content Effect (Question 1a)

Students answered 59.1% of the questions with ED and 60.2% of the questions with MA correctly. The main effect of *content* was not significant (**Table 3**).

Grade-Dependent Content Effect (Everyday vs. Mathematical, Question 1b)

The main effect of grade level was significant (cf. **Table 3**). Over both content conditions, students from grade 6 made significantly more correct inferences than students from grade 4 (67.0% vs. 59.3%, planned contrast $b = 0.51$, $p < 0.01$) and students from grade 4 made more correct inferences than students from grade 2 (59.3% vs. 49.2%, planned contrast $b = 0.54$, $p < 0.01$). The interaction between *content* and *grade level* was not significant (**Table 3**), providing no evidence of a grade-dependent effect of content on conditional reasoning in general (i.e., over all logical forms).

Grade-Independent, Form-Specific Content Effects (Question 2a)

The main and interaction effects referring to logical form (*wording* of minor premise, *type* of conclusion, and their interaction) were significant (**Table 3**). In line with prior results from the everyday contents, a planned contrast analysis showed significantly more correct MP inferences (88.9% of all MP inferences) than MT inferences (69.6%, $b = 1.41$, $p < 0.001$) and significantly more correct MT than AC (41.3%, $b = 1.50$, $p < 0.001$) and DA (48.4%, $b = 1.44$, $p < 0.001$) inferences. Performance of DA and AC inferences did not differ significantly ($b = 0.06$, $p = 1.00$). This indicates that the main effect of *conclusion type* (more correct definite than indefinite responses) was modulated by *wording* of the minor premise only for the definite forms, indicated by lower MT (negative wording) compared to MP (positive wording) performance.

Among the interaction effects with the *content* factor, the interactions with *type* of conclusion (definite vs. indefinite), as well as the three-way interaction of *content*, *type* of conclusion, and *wording* of the minor premise (positive vs. negative) were significant (**Table 3**). For both contents, fewer indefinite correct conclusions were drawn than definite correct conclusions (ED: 36.8% vs. 81.3%, $b = 5.10$, $p < 0.001$; MA: 43.8% vs. 76.4%, $b = 3.61$, $p < 0.001$). Contrary to our expectations, this effect was significantly more pronounced with everyday than with mathematical contents ($b = 1.49$, $p < 0.01$).

A closer analysis indicated that (averaged over all grade levels) there were no significant content-related differences for the two forms with positive wording of the minor premise

TABLE 3 | χ^2 statistics for the fixed main and interaction effects in the final model, in the order of occurrence in the analysis section.

Relates to question	Fixed effect	df	χ^2(df)	*P*	
	Working memory	1	2.24	0.13	
1a	C: content (everyday vs. mathematical)	1	0.92	0.34	
1b	G: grade level	2	24.11	<0.001	***
1b	G × C	2	0.69	0.71	
2a	W: wording of minor premise (positive vs. negative)	1	5.83	0.02	*
2a	T: type of correct conclusion (definite vs. indefinite)	1	94.65	<0.001	***
2a	W × T	1	39.34	<0.001	***
2a	C × W	1	1.12	0.29	
2a	C × T	1	9.92	<0.01	**
2a	C × W × T	1	15.57	<0.001	***
2b	G × W	2	10.64	<0.01	**
2b	G × T	2	0.01	0.99	
2b	G × C × T	2	22.01	<0.001	***

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

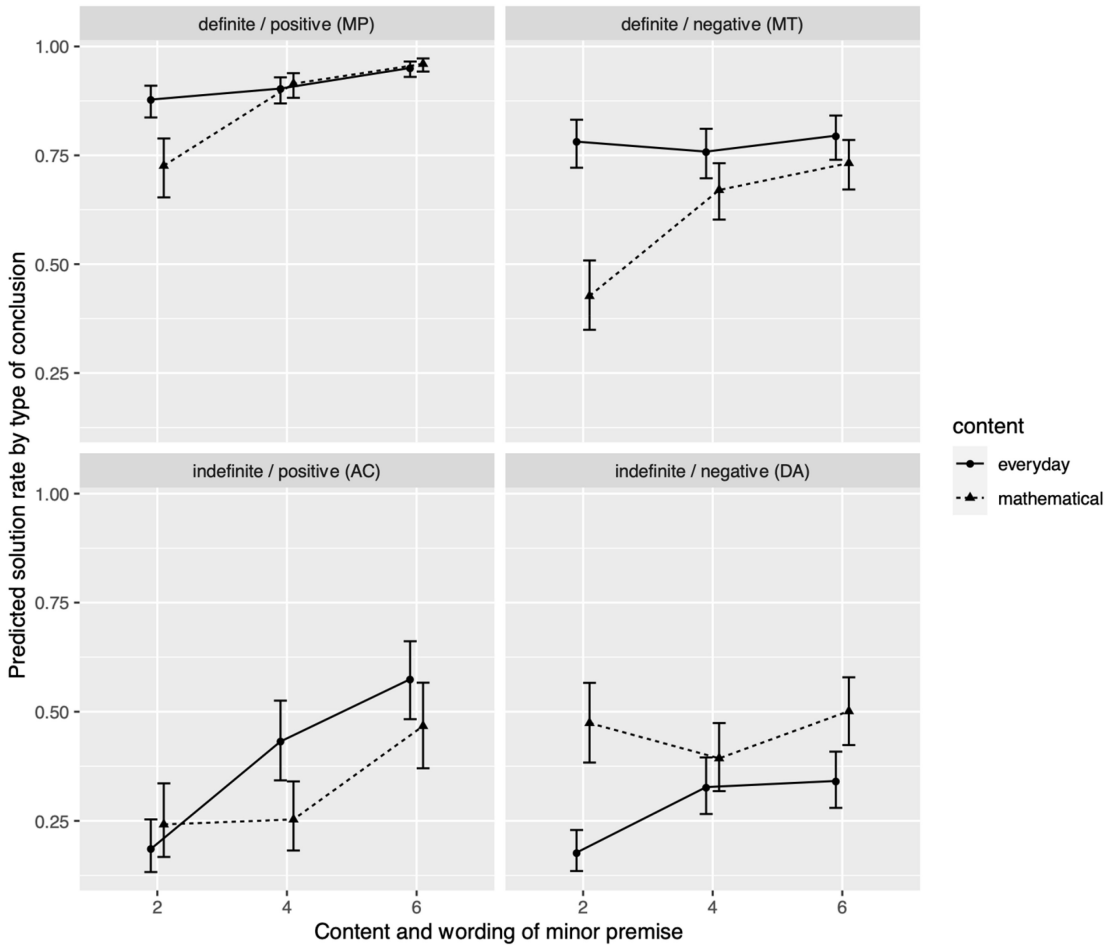


FIGURE 2 | Predicted solution rates and standard error of estimated marginal means (prediction and prediction ± 1 standard error, transformed to the 0-to-1 scale for solution rates) of students' conditional reasoning scores by grade level, logical form, and conditionals' content.

(MP, positive/definite: 89.0% in ED vs. 88.8% in MA, $b = 0.21$, $p = 1.00$; AC, positive/indefinite: 42.0% in ED vs. 40.3% in MA, $b = 0.72$, $p = 1.00$). However, there were marginally more correct MT (negative/definite) inferences with everyday than with mathematical content (73.7% in ED vs. 64.1% in MA, $b = 0.79$, $p = 0.06$) and significantly fewer correct DA (negative/indefinite) inferences with everyday than with mathematical content (31.5% in ED vs. 47.4% in MA, $b = -0.79$, $p < 0.05$). This indicates that the weaker difference between definite and indefinite forms for mathematical, compared to everyday content, was mostly due to the two forms with negative wording. It seems to be mainly caused by better DA (indefinite) and (marginally) lower MT (definite) reasoning with mathematical, compared to everyday content.

Grade- and Form-Dependent Content Effects (Question 2b)

Two significant interactions were observed with a connection to this question. First, as a preliminary result, a significant interaction between *grade level* and *wording* (positive vs. negative) of the minor premise occurred (**Table 3**). Such an effect had not been anticipated because of scarce evidence on the effects of positive versus negative wording of minor premises. A significant performance difference in reasoning with positively worded logical forms over the grade levels could be observed (grade 2: 52.7%, grade 4: 64.6%, grade 6: 74.2%; grade 2 vs. grade 4: $b = 2.92$, $p = 0.02$, grade 4 vs. grade 6: $b = 3.05$, $p < 0.01$). For negatively worded logical forms, only a significant difference between grades 2 and 6 could be found (grade 2: 45.8%, grade 4: 54.0%, grade 6: 59.7%, grade 2 vs. grade 4: $b = 1.36$, $p = 0.44$, grade 4 vs. grade 6: $b = 1.01$, $p = 0.83$, grade 2 vs. grade 6: $b = 2.37$, $p = 0.03$). Even in grade 6, performance on negatively worded logical forms was significantly lower than on positively worded forms ($b = 4.59$, $p < 0.001$), and indeed, differences between grades 2 and 6 were significantly stronger for positively than for negatively worded forms ($b = 3.61$, $p < 0.01$). This speaks for a slower development of reasoning with negatively worded, as compared to positively worded, minor premises.

Moreover, the three-way interaction of *grade level*, *content*, and *conclusion type* (definite vs. indefinite) was significant (**Table 3**). Averaging over all grades, having to draw an indefinite conclusion had already turned out to have a smaller negative impact on reasoning performance with mathematical than with everyday content (see question 1b). Contrast analyses

revealed that, for all grades and both content types, except for mathematical content in grade 2, questions with indefinite correct conclusions lead to lower performance than questions with definite correct conclusions (**Table 4**). Moreover, these contrasts differed significantly between the two contents only in grade 2 (**Table 4**). This indicates that the content-dependent effect of definite versus indefinite conclusions was caused mostly by a stronger difference for everyday, compared to mathematical content, in grade 2.

In particular, this pattern is reflected in better DA reasoning with mathematical than with everyday content in grade 2 ($b = -1.43$, $p < 0.01$). However, this was combined with significantly lower MT reasoning performance with mathematical as compared to everyday content in grade 2 ($b = 1.57$, $p < 0.001$; for MP: $b = 1.00$, $p = 0.11$; for AC: $b = -0.34$, $p = 1.00$).

Exploratory Analyses of Provided Answers

To explain the observed pattern of effects, in particular the better reasoning performance with mathematical compared to the everyday content for indefinite conclusions in grade 2, we analyzed how often students chose the indefinite "uncertain" answer (**Table 5**). We hypothesized that the specific pattern of higher DA and lower MT reasoning could be due to a stronger tendency to give an uncertain response when reasoning about mathematical content, possibly due to difficulties in retrieving or constructing a representation of the problem situation when dealing with the negative forms MT and DA. The amount of indefinite answers increased descriptively from grade 2 to grades 4 and 6 for EDs (**Table 5**). For mathematical content, it was already rather high in grade 2 and remained on this level in grade 4 and grade 6. In grade 2, significantly more indefinite responses were provided with mathematical than with everyday content. In a similar vein, the stronger differences between the two conclusion types by wording of the minor premise (see analyses for question 2a) have to be seen in the context of significantly more indefinite responses on negatively worded questions with mathematical (38.3%) than with everyday content [26.2%, $\chi^2(1) = 23.18$, $p < 0.001$], whereas there was no significant difference on positively worded questions [MA: 22.7% vs. ED: 25.6%, $\chi^2(1) = 1.44$, $p = 0.23$].

TABLE 5 | Frequencies of indefinite ("uncertain") responses by grade level and content, and χ^2 tests for content differences.

Grade	ED	MA	$\chi^2(1)$	*p*
2	18.6%	31.5%	15.7	<0.001
4	28.0%	27.9%	0.0	1.00
6	30.4%	32.2%	0.34	0.56

DISCUSSION

The main goal of this study was to investigate students' reasoning with conditionals about mathematical concepts, which are still emerging during primary school. To this end, reasoning with MAs was contrasted against reasoning with familiar causal premises with EDs. Based on MMT, we started from the assumption that content differences could arise due to general difficulties representing the situations in which the conditionals with MA were embedded, or from specific problems generating alternative models based on these contents. We primarily assumed the latter, because representing the embedding situations for our conditionals did not require specific mathematical knowledge (and we controlled for comprehension of the situations), whereas constructing or retrieving alternatives was strongly contingent on such prior knowledge. Finally, we assumed that both effects could be modulated by the increase in mathematical knowledge during primary school age, leading to more pronounced content effects in earlier compared to later primary school grades. Given the sparse evidence, we did not put forward hypotheses regarding positive or negative wording of the minor premise.

Overall Performance

Considering average performance over all logical forms and grades, more than 50% of students' replies were correct, which is substantially above a guessing probability of 33.3%. This could be taken as evidence in favor of the claim that elementary students do possess early conditional reasoning skills to some extent (e.g., Markovits and Thompson, 2008). In line with previous studies (e.g., Klaczynski and Narasimham, 1998; Markovits and Barrouillet, 2002; Klaczynski et al., 2004; Gauffroy and Barrouillet, 2009), the results indicate that students' performance increases with grade level. Moreover, in line with prior research

TABLE 4 | Solution rates and contrasts between definite and indefinite forms by grade level.

Grade	Content	Solution rates				Definite/indefinite contrast		Content contrast	
		MP	MT	AC	DA	*b*	*p*	*b*	*p*
2	ED	85.9%	70.9%	21.1%	21.9%	6.26	<0.001	4.33	<0.001
2	MA	71.6%	44.8%	30.3%	44.8%	1.93	0.15		
4	ED	86.4%	73.5%	46.2%	34.1%	4.37	<0.001	−0.22	1.00
4	MA	88.6%	67.9%	36.8%	40.6%	4.59	<0.001		
6	ED	93.9%	76.4%	56.5%	37.7%	4.67	<0.001	0.36	1.00
6	MA	97.8%	71.0%	48.1%	54.2%	4.31	<0.001		

The content contrast is the difference of the two definite/indefinite contrasts for the respective grade level.

(e.g., Barrouillet and Lecas, 1999), we found that MP reasoning was easier for elementary school students than MT reasoning, which was in turn easier than AC and DA reasoning, averaging over both contexts and logical forms (Markovits et al., 1996; Markovits, 2000). However, our results did not completely match our predictions made on the basis of MMT.

General Content Effects

We did not find a general effect of content, averaging over all logical forms and grade levels. This indicates that conditionals with mathematical content did not pose general difficulties for conditional reasoning (beyond comprehension of the framing situations, which was controlled by the comprehension questions). It is important to note, however, that the conditionals with mathematical content were not symbolic mathematical statements, but statements about situations, which included a mathematical structure (e.g., the equivalent of different kinds of collections of red and blue gems in gold coins, or the number of windows in a "dwarf house"). We had assumed that mathematical knowledge was not primarily necessary to represent the conditionals, but to construct alternative models (e.g., different collections of red and blue gems with the same overall value). In this sense, the non-significant main effect of the conditionals' content is in line with the rationale of the conditional reasoning tasks applied in this study.

Similar conclusions can be drawn for the interaction of conditionals' content and grade level. Indeed, we did not find evidence of different grade-related differences of overall conditional reasoning between the two contents. This replicates the first results from our own pilot study (Datsogianni et al., 2018). Again, given our assumptions about the necessity of mathematical knowledge in our tasks, this is in line with our expectations. The missing general content difference (everyday vs. mathematical) seems to contradict results from previous studies that the conditionals' content does play a role in conditional reasoning (Markovits and Lortie-Forgues, 2011). However, previous studies (e.g., Markovits and Vachon, 1990; Markovits and Lortie-Forgues, 2011; Markovits, 2014) compared different kinds of relations between antecedent and consequent in everyday contexts (e.g., categorical, causal, or counterfactual conditionals) that are assumed to require different levels of abstraction. Contrary to this, we used causal conditionals with EDs and corresponding conditionals with a mathematical structural mechanism mediating between antecedent and consequent. We assumed that the availability of mathematical knowledge would influence the retrieval of alternative mental models in the mathematical content condition, specifically. That we did not find general content effects seems to indicate, on first sight, that results on early conditional reasoning skills (Markovits et al., 1996) can be transferred from everyday reasoning with familiar causal conditionals to reasoning with mathematical concepts. A detailed analysis of logical forms provided a more differentiated picture.

Definite Versus Indefinite Conclusions

Along our line of reasoning, we had expected a pronounced interaction of conditionals' content with the type of conclusion necessary for a given inference: If mathematical knowledge was primarily necessary to construct alternative models for conditionals with mathematical content, a disadvantage should occur for those inferences that require an indefinite conclusion. Thus, we expected that the difference between performance on tasks that require a definite versus an indefinite conclusion would be more pronounced for mathematical contents. However, what we found was the opposite pattern: reasoning on indefinite forms actually turned out easier for mathematical than for everyday contents, and this effect was particularly pronounced for negatively phrased minor premises (MT vs. DA reasoning). This finding is not in line with our *a priori* predictions based on MMT, and we had no *a priori* explanation for such a result in the context of MMT. It must, however, be seen in the context of descriptively lower MT performance with mathematical, as compared to everyday contents. One reason for this pattern of results might be a stronger tendency for indefinite conclusions ("you cannot say for sure whether. . . or not. . .") when reasoning about mathematical content on some logical forms: Negatively worded premises do not provide a specific situation (e.g., worth 15 gold coins), but only give an indication about what is *not* the case (e.g., not worth 15 gold coins). In contrast to mathematical content, it seems plausible that most students can construct specific models for everyday negation statements such as "not wet" ("dry"). Thus, for mathematical content, students might have problems to retrieve alternative models. The resulting failure to apply reasoning schemata, which they would have applied with everyday content, for conditionals with mathematical content might have weakened students' beliefs about their answers or even lead to an increase in guessing. Indeed, the distribution of responses did not deviate from a uniform (guessing) distribution significantly for mathematical DA and MT items in grade 2. Insecurity or guessing, in turn, could be an explanation for an increased number of indefinite responses when reasoning with mathematical and negatively worded conditionals. This mechanism cannot be investigated with the data at hand, but it could be tested in future research.

Similarly, reasoning on logical forms with indefinite correct conclusions did not turn out to be significantly harder with mathematical than with everyday content for grade 2 reasoners. Again, this has to be seen in the light of a stronger tendency of second graders to give indefinite responses when reasoning about mathematical than for everyday contents. Thus, a similar mechanism as described before might explain this specific effect for grade 2, if we assume that second graders had most problems constructing alternative models to implement relevant reasoning schemata.

We had assumed that problems in alternative generation would lead to more false definite responses on the indefinite forms (AC, DA) in the mathematical context. Our findings do not support this idea. Based on the observed pattern of results, we propose that problems with generating alternatives in the mathematical context may not lead to more incorrect definite answers, but rather to more indefinite answers. Consistent with the idea that mathematical knowledge is important for conditional reasoning, when reasoning with indefinite logical forms about this kind of content, the knowledge about the conditionals' content seemed to modulate the overall positive development of AC and DA reasoning across age

found in prior studies (Janveau-Brennan and Markovits, 1999; Markovits and Barrouillet, 2002). This effect could delay the performance increase over primary school for mathematical content, but also lead to higher solution rates than with everyday contents (for which sound knowledge can be assumed) compared to MAs. Consistent with existing studies (Gauffroy and Barrouillet, 2009; Ricco, 2010; Moshman, 2011; Markovits, 2014), logical form turned out as a key factor in describing conditional reasoning performance, in this case, regarding the contrast between conditionals with mathematical and everyday contents.

The existence of (form-specific) content effects, in any case, supports the assumption that conditional reasoning is sensitive to domain differences at least in early stages of its development, as it was hypothesized in prior work (Chao and Cheng, 2000). The mechanism described above to explain a stronger tendency for indefinite conclusions for mathematical content would indicate that this domain specificity might originate from the fact that acquired reasoning schemata are generally applicable, but still dependent on domain-specific knowledge, unless they develop into more abstract reasoning schemata that work without recourse to domain-specific knowledge. However, given the unexpected pattern of results, this proposed mechanism will have to be investigated in further research. In particular, this result does not contradict the basic assumptions of MMT that the conditional inferences are derived not only from a syntactic analysis of the conditionals (based on knowledge stored in long-term memory) but also from a semantic analysis of the conditionals' contents (Markovits et al., 1998). However, also other accounts of conditional reasoning are discussed in the literature, which could provide alternative explanations for the observed result pattern [e.g., the dual-source model of probabilistic conditional reasoning proposed[1] by Klauer et al. (2010)].

Positive Versus Negatively Worded Minor Premises

Regarding positive and negative wording of the minor premise, we found lower performance and a slower increase of performance on negatively phrased minor premises. Even though this wording–grade interaction was not further qualified by an interaction with the conditionals' content in our analyses, we cannot exclude the existence of such a moderation because of the restricted power of our study for this interaction (see **Supplementary Materials A.3**). If these findings can be sustained with larger samples, this could indicate that the difficulties of negatively phrased minor premises, which have been mentioned in the literature (Schroyens et al., 2001) before, would not differ very strongly between everyday reasoning and the kind of reasoning with mathematical concepts we studied. Moreover, given that negatively phrased forms are investigated less frequently, the mechanisms leading to such a difference can only be hypothesized, at this point. One reason, for example, could be a difficulty representing negations in terms of mental models, which are usually assumed to represent what is *possible* under certain assumptions, not what is *impossible* (Johnson-Laird, 2001). A first indication in this direction is that our findings do not allow us to rule out the application of guessing strategies on negatively worded items in the mathematical context in grade 2 students.

[1]According to this model, inferences are based on two sources of evidence: logical form (decontextualized source of evidence) and prior knowledge about the reasoning context.

Limitations

Our study has to be considered in light of a set of limitations caused by its specific design. First, we used specific tasks to study conditional reasoning with mathematical concepts, which do not reflect deductive reasoning within a mathematical theory. We considered reasoning about mathematical structures embedded in meaningful situations to be more appropriate to study the role of mathematical concepts in conditional reasoning. Extending the results to deductive reasoning in mathematical theories, as it occurs in later years of education, however, is not straightforward. Second, our study has of course limited statistical power to identify small contrasts. A *post hoc* power analysis indicated that only contrast coefficients of up to $b = 1.50$ ($b = 1.00$ for wording of the minor premise) can be identified reliably for main effects in our setting (see **Supplementary Materials A.3**), which is similar to some of the observed contrast values. Insignificant findings cannot be taken as evidence for parallel developments or null effects. On the other hand, the identified differences do offer support for accounts that argue for a role of knowledge about the conditionals' contents in conditional reasoning. Third, given the cross-sectional design, we cannot draw inferences on the individual development of primary school students' conditional reasoning. In particular, the large number of excluded answers due to task comprehension for mathematical content items in grade 2 is an issue here. Beyond the grade-level contrasts investigated in our study, future research should also focus on individual developmental trajectories for both reasoning contents, possibly in interaction with the development of mathematical knowledge and skills. Relatedly, our study focused on the primary school age from grades 2 to 6, which is a key phase for the development of everyday conditional reasoning with causal premises. However, content effects might arise at earlier (e.g., for MP) or at later ages, when AC and DA reasoning become more secure. The latter would also correspond to complaints about problems in conditional reasoning in secondary school students (Küchemann and Hoyles, 2002). Future research could extend the current findings beyond primary school age. As for earlier ages, the availability of the required knowledge about mathematical concepts would have to be taken into account carefully, because these concepts are usually not introduced before grades 1 and 2 of primary school.

Summary

Our results go beyond previous reports on conditional reasoning with everyday concepts and show that even elementary students are able to make valid deductions for some logical forms when reasoning about mathematical concepts. We acknowledge that there are considerable discussions among researchers about students' ability to make conditional deductions, as well as its central importance for scientific reasoning, hypotheses generation and evaluation (Kuhn et al., 1988) and for mathematical thinking (Moshman, 1990;

Markovits and Lortie-Forgues, 2011). Our results are parallel to previous results on other scientific styles (e.g., experimentation: Osterhaus et al., 2015; or probabilistic reasoning: Saffran et al., 2016) that elementary school students are able to engage in correct scientific (deductive, in this case) reasoning in specific conditions. However, our results also underpin findings that these skills may be limited to certain conditions, as MP and MT reasoning in our case, or, for example, certain kinds of covariation data in Saffran et al. (2016). In particular, understanding which factors influence students' scientific reasoning offers starting points for evidence-based support of students' development.

For deductive reasoning, our results provide new perspectives on the role of some knowledge about the concepts involved in the statements used for conditional reasoning. This is in line with findings that students do not use general, abstract reasoning rules at this age (Chao and Cheng, 2000). The proposed mechanism describes how knowledge about the conditional contents and more general conditional reasoning skills could interact and develop over elementary school age. In this account, weak mathematical knowledge might inhibit reasoning in forms (e.g., MT) that not only are at least partially mastered in more familiar contexts according to the literature (e.g., Markovits, 2000; Markovits and Thompson, 2008), but also lead to more correct answers on other logical forms (e.g., DA). Our study provides indications that with increasing familiarity with mathematical concepts in higher grades, performance at least in reasoning with definite forms (MP and MT) on mathematical concepts approaches performance in everyday contexts. All in all, our results are still in line with a model that puts both mathematical knowledge and conditional reasoning strategies as necessary and mutually non-compensating prerequisites of conditional reasoning with mathematical concepts.

If mathematical knowledge is necessary for conditional reasoning with these concepts, it is an open question if this connection can be used in the other direction: Experiencing conditional inferences with mathematical concepts and discussing alternatives as well as other models for the involved conditionals might not only help to increase conditional reasoning skills, but also add to students' knowledge about these concepts. This is in line with current standards (National Governors Association Center for Best Practices, and Council of Chief State School Officers, 2010): argumentation, proof, and reasoning should be incorporated regularly into the mathematics classroom from prekindergarten through grade 12.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Centre of Educational Research and Evaluation of Cyprus Pedagogical Institute and Cyprus Ministry of Education and Culture. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

AD, SU, and BS contributed to the design of the study. AD collected all data. AD, BS, HM, and SU discussed the statistical analysis. AD and SU performed the statistical analysis. AD wrote the first draft of the manuscript. SU revised parts of the manuscript. BS and HM provided feedback to the original manuscript and its revisions. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

This study was funded by Elite Network of Bavaria through the International Doctoral Program REASON in Munich, Germany.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.531640/full#supplementary-material

REFERENCES

Attridge, N., and Inglis, M. (2013). Advanced mathematical study and the development of conditional reasoning skills. *PLoS one* 8:e69399. doi: 10.1371/journal.pone.0069399

Barrouillet, P., and Lecas, J. F. (1999). Mental models in conditional reasoning and working memory. *Think. Reason.* 5, 289–302. doi: 10.1080/135467899393940

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (2018). Parsimonious mixed models. arXiv:1506.04967.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Statist. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Chao, S. J., and Cheng, P. W. (2000). The emergence of inferential rules: the use of pragmatic reasoning schemas by preschoolers. *Cogn. Dev.* 15, 39–62. doi: 10.1016/S0885-2014(00)00018-6

Christoforides, M., Spanoudis, G., and Demetriou, A. (2016). Coping with logical fallacies: a developmental training program for learning to reason. *Child Dev.* 87, 1856–1876. doi: 10.1111/cdev.12557

Crombie, A. C. (1994). *Styles of Scientific Thinking in the European Tradition: The History of Argument and Explanation Especially in the Mathematical and Biomedical Sciences and Arts*, Vol. 1. London: Duckworth.

Cummins, D. (1996b). Evidence of deontic reasoning in 3- and 4-year-old children. *Mem. Cogn.* 24, 823–829. doi: 10.3758/bf03201105

Cummins, D. D. (1995). Naive theories and causal deduction. *Mem. Cogn.* 23, 646–658. doi: 10.3758/bf03197265

Cummins, D. D., Lubart, T., Alksnis, O., and Rist, R. (1991). Conditional reasoning and causation. *Mem. Cogn.* 19, 274–282. doi: 10.3758/bf03211151

Datsogianni, A., Ufer, S., and Sodian, B. (2018). "Elementary students' conditional reasoning skills in mathematical and everyday contexts," in *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics*

Education, Vol. 2, eds E. Bergqvist, M. Österholm, C. Granberg, and L. Sumpter (Umeå: PME), 331–338.

Durand-Guerrier, V. (2003). Which notion of implication is the right one? From logical considerations to a didactic perspective. *Educ. Stud. Math.* 53, 5–34. doi: 10.1023/A:1024661004375

Evans, J. S. B. T., and Over, D. E. (2004). *If*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198525134.001.0001

Foltz, C., Overton, W. F., and Ricco, R. B. (1995). Proof construction: adolescent development from inductive to deductive problem-solving strategies. *J. Exp. Child Psychol.* 59, 179–195. doi: 10.1006/jecp.1995.1008

Gauffroy, C., and Barrouillet, P. (2009). Heuristic and analytic processes in mental models for conditionals: an integrative developmental theory. *Dev. Rev.* 29, 249–282. doi: 10.1016/j.dr.2009.09.002

Handley, S. J., Capon, A., Beveridge, M., Dennis, I., and Evans, J. S. B. (2004). Working memory, inhibitory control and the development of children's reasoning. *Think. Reason.* 10, 175–195. doi: 10.1080/13546780442000051

Holmes, J., and Adams, J. W. (2006). Working memory and children's mathematical skills: implications for mathematical development and mathematics curricula. *Educ. Psychol.* 26, 339–366. doi: 10.1080/01443410500341056

Inglis, M., and Simpson, A. (2008). Conditional inference and advanced mathematical study. *Educ. Stud. Math.* 67, 187–204. doi: 10.1007/s10649-007-9098-9

Inglis, M., and Simpson, A. (2009). Conditional inference and advanced mathematical study: further evidence. *Educ. Stud. Math.* 72, 185–198. doi: 10.1007/s10649-009-9187-z

Janveau-Brennan, G., and Markovits, H. (1999). The development of reasoning with causal conditionals. *Dev. Psychol.* 35:904. doi: 10.1037/0012-1649.35.4.904

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends Cogn. Sci.* 5, 434–442. doi: 10.1016/s1364-6613(00)01751-4

Johnson-Laird, P. N., and Byrne, R. M. (2002). Conditionals: a theory of meaning, pragmatics, and inference. *Psychol. Rev.* 109:646. doi: 10.1037/0033-295x.109.4.646

Kind, P., and Osborne, J. (2017). Styles of scientific reasoning: a cultural rationale for science education? *Sci. Educ.* 101, 8–31. doi: 10.1002/sce.21251

Klaczynski, P. A., and Narasimham, G. (1998). Representations as mediators of adolescent deductive reasoning. *Dev. Psychol.* 34:865. doi: 10.1037/0012-1649.34.5.865

Klaczynski, P. A., Schuneman, M. J., and Daniel, D. B. (2004). Theories of conditional reasoning: a developmental examination of competing hypotheses. *Dev. Psychol.* 40:559. doi: 10.1037/0012-1649.40.4.559

Klauer, K. C., Beller, S., and Hütter, M. (2010). Conditional reasoning in context: a dual-source model of probabilistic inference. *Learn. Mem. Cogn.* 36, 298–323. doi: 10.1037/a0018705

Küchemann, D., and Hoyles, C. (2002). "Students' understanding of a logical implication and its converse," in *Proceedings 26th Conference of the International Group for the Psychology of Mathematics Education*, Vol. 3, ed. A. Cockburn (Norwich: PME), 3–241.

Kuhn, D., Amsel, E., O'Loughlin, M., Schauble, L., Leadbeater, B., and Yotive, W. (1988). *The Development of Scientific Thinking Skills*. Cambridge, MA: Academic Press.

Litkowski, E. C., Duncan, R. J., Logan, J. A., and Purpura, D. J. (2020). When do preschoolers learn specific mathematics skills? Mapping the development of early numeracy knowledge. *J. Exp. Child Psychol.* 195:104846. doi: 10.1016/j.jecp.2020.104846

Markovits, H. (2000). A mental model analysis of young children's conditional reasoning with meaningful premises. *Think. Reason.* 6, 335–347. doi: 10.1080/135467800750038166

Markovits, H. (2014). "Conditional reasoning and semantic memory retrieval," in *Reasoning as Memory*, eds A. Feeney and V. Thompson (New York, NY: Psychology Press), 67–84.

Markovits, H. (2017). In the beginning stages: conditional reasoning with category based and causal premises in 8-to 10-year olds. *Cogn. Dev.* 41, 1–9. doi: 10.1016/j.cogdev.2016.10.002

Markovits, H., and Barrouillet, P. (2002). The development of conditional reasoning: a mental model account. *Dev. Rev.* 22, 5–36. doi: 10.1006/drev.2000.0533

Markovits, H., Fleury, M. L., Quinn, S., and Venet, M. (1998). The development of conditional reasoning and the structure of semantic memory. *Child Dev.* 69, 742–755. doi: 10.1111/j.1467-8624.1998.tb06240.x

Markovits, H., and Lortie-Forgues, H. (2011). Conditional reasoning with false premises facilitates the transition between familiar and abstract reasoning. *Child Dev.* 82, 646–660. doi: 10.1111/j.1467-8624.2010.01526.x

Markovits, H., and Quinn, S. (2002). Efficiency of retrieval correlates with "logical" reasoning from causal conditional premises. *Mem. Cogn.* 30, 696–706. doi: 10.3758/bf03196426

Markovits, H., and Thompson, V. (2008). Different developmental patterns of simple deductive and probabilistic inferential reasoning. *Mem. Cogn.* 36, 1066–1078. doi: 10.3758/mc.36.6.1066

Markovits, H., and Vachon, R. (1990). Conditional reasoning, representation, and level of abstraction. *Dev. Psychol.* 26:942. doi: 10.1037/0012-1649.26.6.942

Markovits, H., Venet, M., Janveau-Brennan, G., Malfait, N., Pion, N., and Vadeboncoeur, I. (1996). Reasoning in young children: fantasy and information retrieval. *Child Dev.* 67, 2857–2872. doi: 10.2307/1131756

Morsanyi, K., Kahl, T., and Rooney, R. (2017). "The link between math and logic in adolescence: The effect of argument form," in *Individual Differences in Judgment and Decision Making From a Developmental Context*, eds M. E. Toplak and J. Weller (Hove: Psychology Press), 166–185.

Morsanyi, K., McCormack, T., and O'Mahony, E. (2018). The link between deductive reasoning and mathematics. *Think. Reason.* 24, 234–257. doi: 10.1080/13546783.2017.1384760

Morsanyi, K., and Szücs, D. (2014). "The link between mathematics and logical reasoning," in *The Routledge International Handbook Of Dyscalculia And Mathematical Learning Difficulties*, ed. S. Chinn (London: Routledge), 101–114.

Moshman, D. (1990). "The development of metalogical understanding," in *Reasoning, Necessity, And Logic: Developmental Perspectives*, ed. W. Overton (Hove: Lawrence Erlbaum Associates), 205–225.

Moshman, D. (2011). *Adolescent Rationality and Development: Cognition, Morality, and Identity*. Milton Park: Taylor & Francis. doi: 10.4324/9780203835111

Nakamichi, K. (2007). Young children's conditional reasoning, working memory, and inhibitory control. *Jap. J. Educ. Psychol.* 55, 347–358. doi: 10.5926/jjep1953.55.3_347

Nakamichi, K. (2011). Age differences in the relationship among conditional inference, working memory and prepotent response inhibition. *Psychologia* 54, 52–66. doi: 10.2117/psysoc.2011.52

National Governors Association Center for Best Practices, and Council of Chief State School Officers (2010). *Common Core State Standards (Mathematics)*. Washington, DC: Council of Chief State School Officers.

Newstead, S. E., Handley, S. J., Harley, C., Wright, H., and Farrelly, D. (2004). Individual differences in deductive reasoning. *Quart. J. Exp. Psychol. Sec. A* 57, 33–60. doi: 10.1080/02724980343000116

Nickerson, R. (2015). *Conditional Reasoning: The Unruly Syntactics, Semantics, Thematics, and Pragmatics of "if"*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780190202996.001.0001

Nunes, T., Bryant, P., Evans, D., Bell, D., Gardner, S., Gardner, A., et al. (2007). The contribution of logical reasoning to the learning of mathematics in primary school. *Br. J. Dev. Psychol.* 25, 147–166. doi: 10.1348/026151006x153127

Osterhaus, C., Koerber, S., and Sodian, B. (2015). Children's understanding of experimental contrast and experimental control: an inventory for primary school. *Front. Learn. Res.* 3:56–94. doi: 10.14786/flr.v3i4.220

Paulus, C. (2009). *Die 'Bücheraufgabe' zur Bestimmung des kulturellen Kapitals bei Grundschülern [The 'book task' as a measure of the cultural capital with primary pupils]*. Available online at: http://hdl.handle.net/20.500.11780/3344 (accessed May 15, 2020).

Ricco, R. B. (2010). "The development of deductive reasoning across the lifespan," in *Biology, Cognition, And Methods Across The Life-Span: Handbook Of Life-Span Development*, Vol. 1, ed. W. F. Overton (Hoboken: Wiley), 391–430. doi: 10.1002/9780470880166.hlsd001012

Robinson, K. M., Price, J. A., and Demyen, B. (2018). Understanding arithmetic concepts: does operation matter? *J. Exp. Child Psychol.* 166, 421–436. doi: 10.1016/j.jecp.2017.09.003

Saffran, A., Barchfeld, P., Sodian, B., and Alibali, M. W. (2016). Children's and adults' interpretation of covariation data: oes symmetry of variables matter? *Dev. Psychol.* 52:1530. doi: 10.1037/dev0000203

Schroyens, W. J., Schaeken, W., and d'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Think. Reason.* 7, 121–172. doi: 10.1080/13546780042000091

Stylianides, A. J., Stylianides, G. J., and Philippou, G. N. (2004). Undergraduate students' understanding of the contraposition equivalence rule in symbolic and verbal contexts. *Educ. Stud. Math.* 55, 133–162. doi: 10.1023/b:educ.0000017671.47700.0b

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Datsogianni, Sodian, Markovits and Ufer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 22 July 2020
doi: 10.3389/fpsyg.2020.01503

Prompting Children's Belief Revision About Balance Through Primary and Secondary Sources of Evidence

*Nicole E. Larsen, Vaunam P. Venkadasalam and Patricia A. Ganea**

Department of Applied Psychology and Human Development, University of Toronto, Toronto, ON, Canada

Prior evidence has shown that children's understanding of balance proceeds through stages. Children go from a stage where they lack a consistent theory (*No Theory*), to becoming *Center Theorists* at around age 6 (believing that all objects balance in their geometric center), to *Mass Theorists* at around age 8, when they begin to consider the distribution of objects' mass. In this study we adapted prior testing paradigms to examine 5-year-olds' understanding of balance and compared children's learning about balance from evidence presented through primary sources (a guided activity) or secondary sources (picture books). Most of the research on young children's understanding of balance has been conducted using a single object, weighted either proportionally (symmetrical object) or disproportionally (asymmetrical object). In this study, instead of using a single object, 5-year-olds (N = 102) were shown 4 pairs of objects, two with the same weight and two with different weight. Children were told to place the objects on a beam where they thought they would balance. We found evidence for an intermediate level of understanding. *Transition Theorists* represent children who have two distinct theories, one for balancing same weight objects, and one for balancing different weight objects, but one of these theories is incorrect. Following the assessment of children's understanding, we compared their learning about balance from evidence that was either presented through primary sources (a guided activity) or secondary sources (picture books). Children learn equally well from both sources of evidence. Findings are discussed in terms of theoretical and practical implications.

Keywords: science learning, belief revision, balance, anomalous evidence, explanations, picture books, guided activity

OPEN ACCESS

Edited by:
Christopher Osterhaus, Ludwig Maximilian University of Munich, Germany

Reviewed by:
Barbara Koslowski, Cornell University, United States
Grace Murray, Kent State University, United States

***Correspondence:**
Patricia A. Ganea
patricia.ganea@utoronto.ca

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *11 March 2020*
Accepted: *05 June 2020*
Published: *22 July 2020*

Citation:
Larsen NE, Venkadasalam VP and Ganea PA (2020) Prompting Children's Belief Revision About Balance Through Primary and Secondary Sources of Evidence. Front. Psychol. 11:1503. doi: 10.3389/fpsyg.2020.01503

INTRODUCTION

Children build naive theories about the world around them and the physical rules that govern it through their daily first-hand observations and experiences (Brewer et al., 1998; Baillargeon, 2002). For some concepts children create beliefs that are counter to valid scientific conceptions. For instance, children may observe a bowling ball and a feather falling and may develop the incorrect idea that heavier objects fall faster than light ones. However, for other concepts, children's conceptions may be correct, but partly incomplete (Clement, 2013). For instance, children learning to balance objects first form the belief that objects balance at their geometric center. This is true

only for objects with an evenly distributed mass and represents a partial understanding of balance principles. Revising prior beliefs can be difficult because naive theories are built on the basis of first-hand observations and can also be driven by cognitive biases (Karmiloff-Smith and Inhelder, 1974; Shtulman, 2017; Kelemen, 2019). Also, just like adults, children easily interpret events that support a theory, but treat counterevidence (i.e., anomalies) as exemptions and isolated cases from their current theory (Karmiloff-Smith and Inhelder, 1974; Chinn and Brewer, 1993). Even when they consider the counterevidence, they tend to create a new theory to account for it, independent of their existing theory (Karmiloff-Smith and Inhelder, 1974). The current study examines children's belief revision about balance relations. Specifically, we investigated how children's beliefs about balance are impacted when the evidence is presented through primary sources (guided activity) or secondary sources (picture books). We also developed a new framework for characterizing children's theories about balance that accounts for intermediate levels of understanding.

Balance Literature

Beginning with the pioneering work of Karmiloff-Smith and Inhelder (1974), research has focused on three theoretical phases that children go through as they develop an understanding of balance. Children younger than six generally balance both symmetrical and asymmetrical blocks using a system of trial and error, with no clear reasoning behind their choices (Karmiloff-Smith and Inhelder, 1974), and so they are usually considered in the "No Theory" stage. Around the age of six, children become "Center Theorists" and begin to balance all blocks using the geometric center of an object as their reference point (Karmiloff-Smith and Inhelder, 1974). That is, when given a single block to balance, they place the geometric center on the fulcrum, as opposed to the center of mass, regardless of whether they are balancing symmetric or asymmetric blocks. Around the age of seven or eight, children begin to consider the distribution of weight, thereby using the center of mass as the balance point (Karmiloff-Smith and Inhelder, 1974). "Mass Theorists" can correctly balance both symmetrical and asymmetrical blocks.

Research by Siegler (1976) has also shown that children use four rules as they build an accurate understanding of the concept of balance. Children go from focusing on the relative weight on the two sides of the fulcrum (Rule 1), to focusing on the distance from the fulcrum if there are equal weights on both sides, but only on the relative weight when the weights are unequal (Rule 2), to an understanding based on both weight and distance for equal and unequal weights (Rule 3), and finally to an understanding based on computing torques on both sides (Rule 4). In contrast to the task used by Karmiloff-Smith and Inhelder (1974), in which children had to balance symmetrical and asymmetrical blocks, Siegler (1976) showed children a scale that had either equal and unequal weights placed at various distances from the fulcrum and asked them to predict whether the scale would balance. In this task, 4 and 5-year-old children used weight to inform their predictions (Siegler and Chen, 1998), indicating the presence of an incipient theory of balance.

Though Siegler's balance rules have expanded on the theoretical phases put forth by Karmiloff-Smith and Inhelder, recent developmental research investigating young children's understanding of balance relations has primarily used the placement task developed by the latter (e.g., Bonawitz et al., 2012; Ganea et al., 2017). However, because this task uses a single block, results map either onto Center Theorists (Rule 1 of Siegler's model) or Mass Theorists (Rule 3), without allowing for consideration of an intermediate phase (Rule 2). In this research we combined both tasks to examine whether 5-year-old children hold a hybrid theory for balancing objects. In addition, we aimed to determine whether children's belief revision is influenced by how the evidence is presented.

Evidence Evaluation

Our ability to process and make judgments about evidence is highly influenced by prior beliefs, insofar as we are inclined to draw conclusions that are in line with these beliefs rather than to change them (e.g., Kuhn et al., 1988; Dunbar and Klahr, 1989; Kuhn, 1989; Schauble, 1990; Chinn and Brewer, 1993; Koslowski, 1996; Zimmerman and Klahr, 2018). The process of belief revision can be particularly difficult for children, as they need to accommodate not only the pull of naïve prior beliefs, but the cognitive demands of evidence evaluation as well (Tolmie et al., 2016). Evidence evaluation requires children to be able to track, connect, and remember different instances of similar phenomena; to understand that variables operate in a consistent fashion across said phenomena; and to connect these instances to their theories. Given these high demands on cognitive load, children may fail to engage in belief revision due to the complexity of the task, or because the evidence is not salient enough (Koerber et al., 2017). Demands on cognitive load also come from having to consider multiple variables in the interpretation of anomalous evidence, as is the case for balance (Siegler, 1976). Given these inherent difficulties, an important consideration is how to support children's ability to learn from evidence.

One way to improve learning from anomalous evidence is to provide a plausible explanation that can explain the evidence. An explanation can reduce the bias toward maintaining prior beliefs in the face of counterevidence. In the absence of such an explanation, people are more likely to cling to their prior beliefs (Klahr and Dunbar, 1988; Kuhn, 1989; Chinn and Brewer, 1993). That is, people often consider an inadequate theory as preferable to the absence of a theory. Therefore, combining explanations with anomalous evidence may be the most effective way to promote theory change. There is evidence that children find claims more believable when they can generate, or are provided with, a possible mechanism to explain anomalies (Sandoval et al., 2014). Moreover, children are better able to revise prior beliefs when the evidence is linked to their current beliefs directly (Hardy et al., 2006) and when alternative theories are given prior to the counterevidence (Renken and Nunez, 2010, 2013; Ganea et al., 2020).

Another teaching strategy for increasing children's learning in domains where they have misconceptions involves the use of

multiple examples or analogies (Vosniadou, 1989; Vosniadou and Skopeliti, 2019). The use of analogies through the comparison of multiple examples can lead one to infer the deep abstract features that characterize a concept (Gentner, 1983, 1989) and allows for better visualization of scientific explanations (Vosniadou, 1989; Brown, 1993). In work with third-, fifth-, and sixth-grade children, as well as college students, Vosniadou and Skopeliti (2019) found that text with relevant analogical examples can reduce invalid inferences and misconceptions in understanding of the day/night cycle and seasons. However, work by Brown (1992) found that examples and analogies alone are ineffective when the target problem is in a domain where students hold misconceptions. This may be because students' pre-existing conceptions interfere with the extraction of an abstract schema that accounts for the multiple examples. An effective way to improve this schema abstraction is by adding explanations, which help with the development of analogical connections (Brown, 1992). In this research, we use both examples and explanations to examine children's learning about balance relations.

Sources of Evidence

How evidence is presented to children can also have an impact on their learning. Evidence could be experienced either directly (primary evidence) or indirectly (secondary sources). In science, primary evidence refers to data that is derived from direct experience with an event or object (Sandoval et al., 2014). By extension, learning through primary sources of evidence involves children conducting or witnessing an experiment firsthand. For many concepts, however, primary evidence is inaccessible to the majority of people. Indeed, people are more often forced to consider and evaluate claims without access to such primary evidence, having instead to rely on secondary sources of evidence (Sandoval et al., 2014). Secondary sources of evidence provide an account of primary source evidence through various mediums such as books, websites, or verbal testimony.

Sandoval and Çam (2011) found that third and fourth grade students preferred primary data over other sources, citing the improved credibility of this evidence. Further, second grade students who were asked to evaluate a claim mainly used the primary evidence available to them, ignoring the secondary information (Koerber et al., 2017). Nevertheless, a related body of research indicates that young children have the capacity to learn from secondary sources of evidence, such as another person's verbal testimony about an event or concept (Harris and Koenig, 2006; Harris and Corriveau, 2014). Specifically, with respect to balance principles, research so far has used both primary evidence (i.e., direct experience of anomalous evidence) (Bonawitz et al., 2012) and secondary evidence (i.e., indirect experience of anomalous evidence through a picture book) (Ganea et al., 2017) showing that children can learn from both. In this study we directly compare primary and secondary sources of evidence to examine their relative effectiveness in inducing belief revision about balance in 5-year-old children.

Current Study

The current study directly compares primary and secondary sources of evidence to determine whether or not there are differences in children's ability to learn from them. Each source contained both anomalous evidence and mechanistic explanations. We presented the anomalous evidence through either guided activities (Primary Evidence condition), or through examples described and illustrated in picture books (Secondary Evidence condition). Mechanistic explanations were delivered either verbally by the experimenter during the activity (Primary Evidence condition) or included in the information in the picture books (Secondary Evidence condition). We expected that both conditions will improve from pre- to post-test, but we predicted that the primary evidence condition will see more improvement, based on prior research that children prefer and consider primary evidence to be more credible than secondary evidence (Sandoval and Çam, 2011; Koerber et al., 2017).

We developed our pre- and post-test to combine elements from the tasks of Karmiloff-Smith and Inhelder (1974) and Siegler (1976). First, we asked participants to place four different object pairs (two each of equal and unequal weights) on a beam so that the beam would stay perfectly balanced in the air. This placement element is modeled after Karmiloff-Smith and Inhelder's task, however, we used separate weights to simplify the task and allow for a greater variety of answers. In addition, after children made each placement, we asked them to explain their choices. The use of separate weights and explanations was drawn from Siegler's task. By combining aspects of both tasks for our test phase, we aimed to explore the possibility of intermediate levels of balance understanding.

We examined 5-year-olds's understanding because this is an age for which prior research has provided inconsistent evidence. Five-year-olds were categorized as lacking a theory of balance when tested with the placement task (e.g., Bonawitz et al., 2012; Ganea et al., 2017), but as being able to understand weight as a relevant variable for balance when tested with the prediction task (Siegler and Chen, 1998). Further, Siegler and Chen (1998) found that over the course of several trials, 5-year-old children began to accurately predict balance relationships that were affected by distance. Although children were not able to explain their predictions, they reduced their references to weight. In contrast, 4-year-olds did not improve their predictions and continued to reference weight in their explanations (Siegler and Chen, 1998). Further work by Pine et al. (1999) also found improvement in children's understanding of how to balance asymmetrical blocks in their sample of 5-to-7-year-olds, using a combination of guidance and explanations. This suggests that with support and appropriate evidence, 5-year-olds may be able to move beyond a reliance on weight to understand balance and incorporate distance as another causal dimension.

MATERIALS AND METHODS

Participants

One hundred and two 5-year-olds (M = 5.49; range: 5.02—5.99, 51 females, 51 males) participated in this study. Twenty-two additional children were excluded because they had a perfect score on the pretest (n = 15), had a language barrier (n = 4), or due to parental interference (n = 1) or experimenter error (n = 2).

Children were randomly assigned to one of three conditions: Primary Evidence (M_{age} = 5.51), Secondary Evidence (M_{age} = 5.49), and Control (M_{age} = 5.49). Each condition had 34 children. For our evidentiary conditions, we developed two different hands-on activities and two different books designed to teach children about balance. The use of two different tasks and books allowed us to ensure that children's learning is not tied to a specific type of stimulus. Within each condition, children completed *one* activity or read *one* book. The books and activities were matched in the presentation of anomalous evidence and the amount of conceptual information provided to children. The control condition differed in that children either read an unrelated book or were offered the materials from the hands-on activities to explore independently.

Participants were recruited from a database of families who have expressed interest in participating in research. Children were individually tested by a female experimenter in a quiet room at a University laboratory. The sample of children came from a diverse background, including Asian (13%), Aboriginal (1%), Black (3%), White (47%), and Mixed Race (30%) children. An additional 6% of families declined to disclose ethnicity information. Of the families who disclosed this information the majority of children came from middle- and upper-class families. However, 6% of families declined to disclose this information. The mode parental education level was a Bachelor's Degree, with 5% of families declining to disclose the education for either parent.

Materials

Interventions

In the experimental conditions, children were shown one example of how to balance objects with the same weight, and three examples of how to balance objects with different weight from each source of evidence.

Primary evidence

In the first activity (Placement), children were provided with a beam with different hooks along the beam and different weights that could hang on the hooks. Children hung different weights along the beam to try and balance it (see **Figure 1**). For the second activity (Prediction), children were given a beam with sides that slid closer to and further from the fulcrum. They were also provided with toy crabs, of the same weight, and marbles, jars, and toys of different weights (see **Figure 2**). For this activity children were also given a worksheet to record their predictions and results (the activity scripts can be found in **Supplementary Appendix A**).

Secondary evidence

We wrote and illustrated two texts. The first text, a narrative book, had two young children playing at the park on a seesaw. In the book the children balanced one pair of same weight objects, and three pairs of different weight objects on the seesaw. The second book had the same examples but presented in a straight-forward information text and used photographs instead of pictures (illustrations of the narrative book can be found in **Supplementary Appendix B**; the text from the non-fiction book can be found in **Supplementary Appendix B**, the photographs from the non-fiction book are not included due to reproduction permissions).

Test Phase

As a measure of learning, pre- and post-tests were administered. The materials for each test phase included a stand and a beam with four velcro pieces on it. These velcro pieces were placed with two on either side of the middle, one close to the middle and one further from the middle. We also used four pairs of objects in each test phase, for a total of eight object pairs. All of the objects had pieces of velcro on the bottom, corresponding to the velcro on the beam. Within each set of four pairs, two pairs were the same size and weight, and two pairs were different sizes and weights (see **Figure 3**). For each test phase children received a different object set, but the order in which children received these sets were counterbalanced.

Procedure

This study used a between-subjects design. Each condition followed the same four-phase protocol, where children were randomly assigned to one of three conditions: Primary Evidence, Secondary Evidence, or Control (see **Figure 4**). Within each condition, children were randomly assigned to complete one activity or read one book. The entire session was video-recorded and lasted approximately 15–20 min.

Receptive Vocabulary

To ensure that children were able to understand the tasks and the conceptual information provided in the intervention, we administered the Toolbox Picture Vocabulary Test (TPVT). The TPVT measures receptive vocabulary using an adaptive computerized format (National Institutes of Health, 2015).

Interventions

Primary evidence

In the Primary Evidence condition, children completed one of two different activities: Placement Activity and Prediction Activity.

In the *Placement Activity*, children were provided with a balance beam and several weights. On either side of the beam's fulcrum there were 10 hooks on which children could hang the weights. Each weight was the same, but more than one weight could be hung on the hook (**Figure 1**). The experimenter first hung a single weight on the 5th hook and asked the child to hang a single weight to make the beam balance. Once successful, the child was given two weights. The experimenter then hung a single weight on the 8th hook and asked the child to hang their two weights together to make the beam balance. Next, the experimenter took two weights and hung them on the 3rd hook and asked the child to hang a single weight to make the beam balance. Finally, the child was given two weights and allowed to hang them on any hook from five to one (inclusive). Together, the child and experimenter decided where to hang the experimenter's single weight to balance the beam. In each example, the child kept experimenting until they were successful in their placements. The

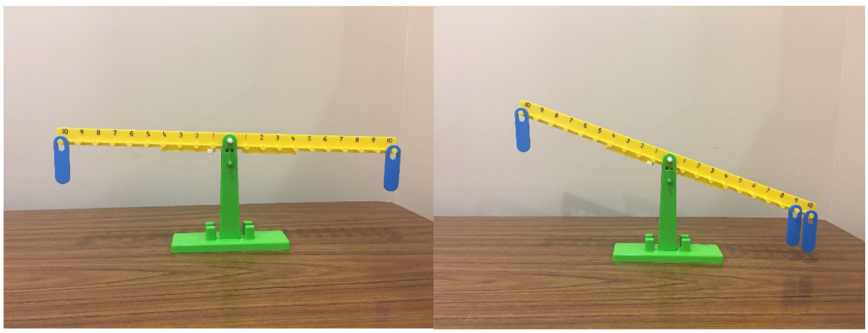


FIGURE 1 | Stimuli for placement activity. Examples of balanced beam with same number of weights on same hook and unbalanced beam with more weights on one side.

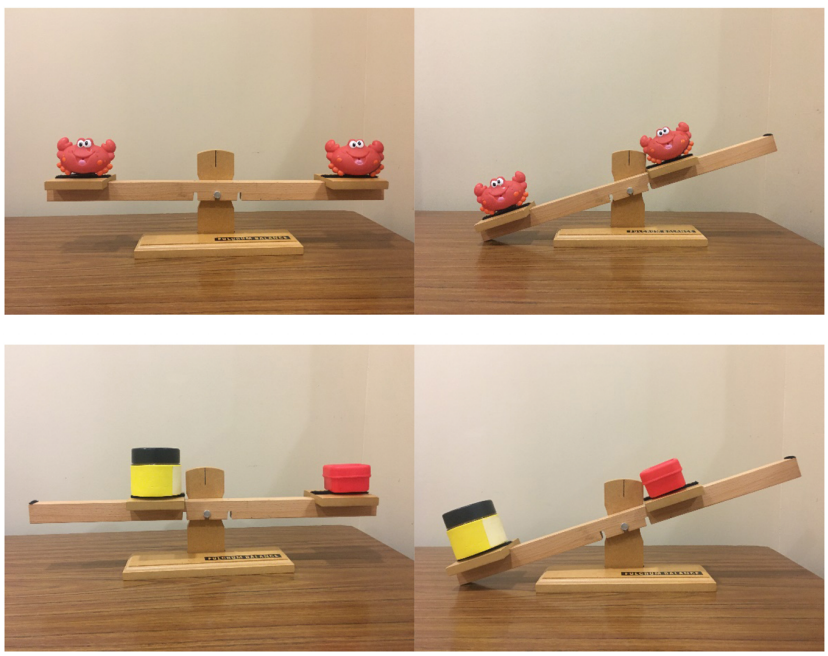


FIGURE 2 | Stimuli for prediction activity. Examples of a balanced beam with same weight objects (crabs) the same distance apart (top left image) and differently weight objects (boxes) with the heavier box closer to the middle (lower left image). Examples of an unbalanced beam with same weight objects (crabs) at different distances (top right image) and differently weight objects (boxes) with the heavier object further from the middle (lower right image).

FIGURE 3 | Stimuli for test phase. Correct placement for different weight objects (blue blocks) and same weight objects (red blocks) shown. Note that same weight objects could also both be placed close to the middle.

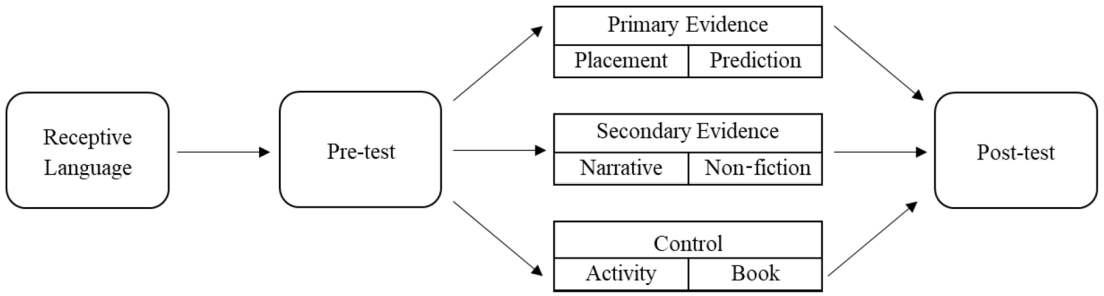


FIGURE 4 | Schematic of study design.

experimenter highlighted that the same weight objects balanced when they were on the same number, and the different weight objects balanced when the heavier ones were closer to the fulcrum and the lighter one was further away.

In the *Prediction Activity* children were provided with a different balance beam. This beam had sliding end seats on it, so that objects could be placed on these seats and then moved closer to or further from the fulcrum (**Figure 2**). Children were also provided with a worksheet, and told they were going to make predictions about how objects balance on a beam and they would record their predictions and the results on this sheet, just like a scientist does. Children were first given two toy crabs with the same weight. They put these crabs on the seats of the beam and moved the seats until they were able to balance the beam. They were then provided with two marbles that were different weights. They again placed the marbles on the seats and moved them until they were able to balance. This was repeated with two more object pairs that had different weights: jars and toys. As with the Placement Activity, the experimenter highlighted that the same weight objects balanced when they were on the same number, and the different weight objects balanced when the heavier ones were closer to the fulcrum and the lighter one was further away.

For both activities, the intervention ended with the experimenter providing an explanation about why the objects balanced at different places on the beam. The experimenter explained that objects that have the same weight have the same force on the beam, and therefore they balance the same distance away from the fulcrum. In contrast, when objects have different weight, the heavier object has more force than the lighter one, so the heavier object must be placed closer to the fulcrum, and the lighter object must be placed further away, to compensate for this difference in force.

Secondary evidence

In the Secondary Evidence condition, children were read one of two different picture books, a Narrative Book or a Non-fiction Book (see **Supplementary Appendix B**).

The *Narrative Book* told a story of two children, Alice and Luke, playing on a seesaw. As they are playing Luke wonders why he goes up into the air before Alice. Using experiments with buckets of sand, Alice first demonstrates to Luke how two objects that have the same weight need to be the same distance away from the fulcrum in order to balance. She then uses the buckets to demonstrate that once they have different weights, the heavier bucket needs to be closer to the fulcrum in order to balance. She also demonstrates this with rocks of different weight, and finally with the children themselves. At the end of the book Alice provides Luke with an explanation about why the objects balanced at different places on the beam. This explanation was matched with the explanation the experimenter provided to children who completed the primary evidence activities.

The *Non-fiction Book* matched the *Narrative Book* in terms of the examples and information provided to children. The only difference was that information was presented in an expository format as opposed to embedded in a plot.

Control

In the Control condition, half the children completed a hands-on activity (Activity Control) and the other half were read a picture book (Book Control).

Activity control

In the Activity Control children were provided with all the same materials as the children who completed the Prediction Activity. They were shown how the beam worked and how the seats moved, but were given no further direction with what to do with the beam or the objects, nor any feedback or information about how or why objects balance. The goal of this control was to ensure that mere exposure to the beam was not sufficient for children to learn about balance.

Book control

In the Book Control children were read an unrelated picture book about plants (The Tiny Seed, by Eric Carle). The goal of this control was to ensure that exposure to the pre-test was not sufficient for children to make gains in their understanding on the post-test, due to the parallel structure of these tests of learning.

Test Phase

Pre- and post-tests were administered using the same procedure but with different object sets. Each test phase consisted of four trials, with each trial using one set of four pairs of objects. Each set of four object pairs consisted of two pairs with the same size and weight, and two pairs with different sizes and different weights. In the different pairs, the bigger object was also always the heavier object. Object sets were counterbalanced between test phases, but due to the possibility of differences in response

patterns, within each set the pairs were presented in the same order to participants (same weight, different weight, same weight, and different weight).

During both test phases, participants were first shown a stand and a beam. The experimenter oriented the children to these materials by saying "When I put my beam on my stand like this, it's like a seesaw or a scale. It can stay perfectly balanced in the air, or one side can go up and one side can go down." This served both to introduce the beam to children and highlight the mechanics to children, as well as connect the beam to the intervention phase with the term seesaw (from the books) or scale (from the activities). After this introduction, the experimenter told children "Now, I'm going to show you some blocks and ask you some questions about where we can put the blocks on the beam so that it stays perfectly balanced with both sides in the air." The experimenter also explained to children that they had to put one block on either side of the middle, and demonstrated that on each side, blocks could be placed on two possible velcro strips: one that was closer to the middle, and one that was further from the middle.

Following this orientation children were shown the object pairs one at a time for a total of four trials. For each object pair, the experimenter highlighted that they were either the same size and weight, or different sizes and different weights. Children were asked to place each object pair on the beam so that the beam would stay perfectly balanced in the air. After children placed the objects on the beam, the experimenter placed the beam on the stand, holding it in place so that it would balance regardless of where children had placed the blocks. The experimenter then confirmed with the child "It will balance like this?" If the child answered in the affirmative, the experimenter followed this question with "Why do you think it will balance like this?" Children's explanations were recorded. If the child did not agree it would balance, they were asked to move the blocks to a space where it would balance, and then were asked to provide an explanation for their placements. The experimenter did not let go of the beam, and therefore children could not see if their placement was correct. Children received neutral feedback ("Thank you") after answering each test question.

Coding

Children's placement of the blocks at pre-test and post-test were coded by research assistants using video recordings of the sessions. A visual depiction of the correct placement of the blocks can be found in **Figure 3**. To fit with analyses from previous research, based on these placements, children were categorized into "center theorists" or "mass theorists" or "no theory." Children who had correct placement of all blocks were coded as "mass theorists," and children with completely random placement were coded as "no theorists." Due to the adapted design of our task, we were also able to identify an additional category. We split the center theorists into two groups: "traditional center theorists" and "transition theorists," Traditional center theorists were those that placed all blocks equal distance from the center, regardless of their weight. Transition theorists were those who had separate theories for same weight object pairs and different weight object pairs, though one theory was incorrect; they either placed different weight pairs both close to the center, and same weight object pairs both far from the center (or vice versa), or they placed same weight object pairs correctly, and different weight object pairs with the light object closer to the fulcrum and the heavier object further away. These children were considered Transition as they understood that same weight and different weight objects would need to be balanced differently, but they did not yet have full understanding of the mass theory. This aligns with the rules proposed by Siegler (1976), which specified that children would focus attention on different variables for same weight and different weight object pairs (Rule 2), prior to moving into an understanding of the interplay between distance and weight (Rule 3, or Mass Theorists). Children who placed the same weight pairs correctly and one different weight pair correctly were also coded as transitional theorists.

We also analyzed the justifications children gave when asked to explain their choices for the placements. The explanations were used as a secondary measure of learning and coded on a scale from 0 to 2. A breakdown of the coding scheme for explanations can be found in **Table 1**.

Explanation scores were summed across the four trials, such that scores ranged from 0 to 8 for each test phase. Two research assistants coded 100% of the children's responses. The coders were blind to the hypotheses, test phase and condition of the study. There was high inter-rater reliability determined by Cohen's $\kappa = 0.82$, $p < 0.001$, an 89% agreement rate. The two coders resolved disagreements through discussion.

RESULTS

We conducted two main analyses of the data. First, we examined how children's theories changed from pre- to post-test. Second, we examined children's belief revision as a function of evidence sources from pre- to post-test. Preliminary analyses indicated the two types of books and activities used did not influence the post-test measures. The classification of children as theorists at post-test did not differ by book type, $\chi^2(3, N = 34) = 2.62$,

TABLE 1 | Coding scheme for children's explanations of their balance placement predictions at pre- and post-test.

Score	Placement	Reason	Example
2	Correct	Referencing distance from fulcrum.	Same weight objects: Because they are both far away from the middle. Different weight objects: Because the heavy one it close to the middle.
1	Correct	Referencing distance from fulcrum in conjunction with incorrect information. Reasons that refer to object weight but not distance.	Because they are both far away from the middle and squishy. Because one is heavy and one is light.
0	Correct	Referencing irrelevant information as a justification.	Because they are both square shaped.
0	Incorrect		

p = 0.46 or activity type, $\chi^2(3, N = 34) = 1.44$, $p = 0.61$. Similarly, explanation scores at post-test did not differ by book type, $t(32) = 0.82$, $p = 0.42$, or activity type, $t(32) = 30.29$, $p = 0.92$. As a result we collapsed both types of books and activities in the following analyses.

Theories

Table 2 shows the number of participants categorized as a type of theorist at pre- and post-test across conditions. Chi-Square goodness-of-fit tests were conducted both at pre-test and post-test to examine whether the observed frequency of theorists was different than chance (all 4 categories were treated as equal, chance level = 25.5, 102/4). The observed frequency of theorists differed from chance both before, $\chi^2(3, N = 102) = 9.92$, $p = 0.02$, and after, $\chi^2(3, N = 102) = 41.84$, $p < 0.001$, the intervention. **Table 2** shows that at pre-test 26% of children (26/102) were categorized as no theory, 28% (28/102) as traditional theorists and 34% (35/102) as Transition theorists, but only 13% (13/102) were mass theorists. In contrast, at post-test a smaller percentage of children were categorized as having no theory (17%, 17/102) or traditional theorists (7%, 7/102), and a greater percentage of children were Transition theorists (50%, 51/102) or mass theorists (27%, 27/102).

A multiway frequency analysis (MFA) was used to investigate the relative importance of pre-test, post-test and condition in predicting the expected frequencies, while maintaining adequate fit between expected and observed frequencies. MFA is an extension of the chi-square goodness-of-fit technique, where the goal is to create a model that accounts for the observed and expected frequencies being equal (a non-significant chi-square statistic) while using the fewest variables as possible in the model (Agresti, 2002; Howell, 2007; Tabachnick and Fidell, 2012). MFA is a non-parametric statistical procedure for discrete variables with two or more levels. This type of analysis was selected to accommodate the frequency nature of the dependent variable, and the presence of within-subject factor (pre- and post-test scores) while allowing for the analysis of main effects and interactions (Vokey, 2003).

We generated three preliminary crosstabs comparing pre-test to post-test and each test phase separately by condition to ensure that each cell had sufficient frequencies to conduct an MFA. The crosstabs indicated that no cell had an expected frequency of 0, but 14 of the 40 cells (35%) had frequencies less than 5. Collapsing categories is one way to deal with this violation that no more than 20% of the cells should have expected frequencies below 5 in order to retain appropriate power (Agresti, 2002; Tabachnick and Fidell, 2012). We decided to retain all levels of variables in the tested model, despite the potential loss of statistical power (Wickens, 1989), because all categories were of interest. A *post hoc* power analysis was conducted, revealing that for a calculated effect size $\omega = 0.39$, power was sufficient for this study $(1 - \beta) = 0.80$.

Using a backward hierarchical approach to model building, a saturated log-linear model with all one-way, two-way, and three-way interactions was examined. The likelihood-ratio (LR) χ^2 values for the overall effects found that the two-way interaction between Condition × Post-theories and the one-way effects of Pre- and Post-theories achieved significance (**Table 3**). Backward hierarchical solution statistics identified a model with the two-way Condition × Post-theories interaction and one-way Pre-theories main effect as best fitting, LR $\chi^2(df = 33) = 33.32$, $p = 0.45$. The non-significant LR χ^2 value means that the model has the smallest number of effects that yields an adequate fit between the expected and observed frequencies. A custom model using all significant effects was also assessed and found to be not significant, LR $\chi^2(df = 33) = 31.74$, $p = 0.53$. As both tests were non-significant, we can conclude that the associations dropped from the saturated model were not needed to explain the distribution of frequency data. A visual inspection of plotted residuals confirmed that the observed standardized residuals were acceptably close to those that were expected.

The two-way association between condition and post-theories significantly predicted the expected frequencies. This fits our hypothesis, as we expected that the experimental conditions will

TABLE 2 | Frequencies of children who held a certain balance theory as a function of test-phase and condition.

Condition	Pre-theory	Post-theory				Total
		No theory	Traditional	Transition	Mass	
Primary evidence	No theory	1	0	5	1	7
	Traditional	1	1	7	3	12
	Transition	1	0	4	5	10
	Mass	0	0	2	3	5
	Total	**3**	**1**	**18**	**12**	**34**
Secondary evidence	No theory	1	1	5	1	8
	Traditional	1	1	2	3	7
	Transition	1	0	9	7	17
	Mass	0	0	1	1	2
	Total	**3**	**2**	**17**	**12**	**34**
Control	No theory	4	1	5	1	11
	Traditional	2	1	6	0	9
	Transition	4	2	2	0	8
	Mass	1	0	3	2	6
	Total	**11**	**4**	**16**	**3**	**34**
Total	No theory	6	2	15	3	26
	Traditional	4	3	15	12	28
	Transition	6	2	15	12	35
	Mass	1	0	6	6	13
	Total	**17**	**7**	**51**	**27**	**102**

TABLE 3 | Multiway frequency analysis K-way effects and partial associations.

Effect	Partial χ^2	df	p-value	Number of iterations
Condition	0.00	2	>0.99	2
Pre-test	10.90*	3	0.01	2
Post-test	41.90*	3	<0.001	2
Condition × Pre-test	9.90	6	0.13	2
Condition × Post-test	17.95*	6	0.006	4
Pre-test × Post-test	11.66	9	0.23	4

*$p < 0.05$.

change children's theories compared to the control conditions. From **Table 2**, we see 9 and 11 children from the primary and secondary conditions made Mass Theorist predictions after the intervention, compared to only one child in the control condition. We also observed a greater number of children in the control condition retaining their theory or reverting back to random predictions. The presence of one-way effects suggested the independence of the variables but should not be interpreted if nested in higher order associations (Agresti, 2002). Therefore, we focused exclusively on the pre-theories which was an important predictor of the expected frequencies above and beyond the other effects. The majority of children shifted to a higher ranked theory with only 19 out of 102 children regressing. At post-test, most children were categorized as transition ($n = 51$) and mass ($n = 27$) theorists. Further, most children who were categorized as mass theorists at post-test were traditional and transition theorists at pre-test. This indicates that children's pre-theories affected whether children revised or retained their current theory after the intervention.

Explanations

Preliminary analyses for explanations showed no effect of age, gender or TPVT, therefore these variables were not considered in the following analyses. A one-way ANOVA also revealed that pre-test explanation scores were similar across all three conditions at baseline, $F(2,99) = 1.50$, $p = 0.23$, $\eta_p^2 = 0.03$.

A 2 (test phase) × 3 (condition) mixed-measures ANOVA showed a significant main effect of test phase, $F(1,99) = 78.47$, $p < 0.001$, $\eta_p^2 = 0.44$. Pairwise comparisons with Bonferroni corrections indicated that the mean difference scores at post-test were significantly higher than scores at pre-test ($p < 0.001$, 95% CI = 1.11, 1.80). There was no main effect of condition, $F(2, 99) = 1.52$, $p = 0.220$, $\eta_p^2 = 0.03$. However, there was a significant interaction between test phase and condition, $F(2,99) = 13.24$, $p < 0.001$, $\eta_p^2 = 0.21$. We used estimated marginal means with Bonferroni corrections to determine the nature of this interaction (see **Figure 5**). There was a significant change in children's explanation scores from pre-test to post-test in the primary ($p < 0.001$, 95% CI = 1.54, 2.70) and secondary ($p < 0.001$, 95% CI = 1.48, 2.63) evidence interventions but no significant change for the control condition ($p = 0.38$, 95% CI = −0.31, 0.84). In terms of condition differences at post-test, children in the primary ($p < 0.001$, 95% CI = 0.52, 2.48) and secondary ($p = 0.02$, 95% CI = 0.17, 2.13) evidence conditions had explanations that were significantly higher than those of children in the control condition. There were no differences between children's explanations in the experimental evidence conditions at post-test ($p > 0.99$, 95% CI = −1.33, 0.63).

DISCUSSION

This study evaluated the impact of primary and secondary sources of evidence on children's theories about balance relations. There were two main findings in this research. First, children's prior theories played an important role in determining whether children maintained or modified their beliefs about balance in response to the intervention. Second, children were able to learn that distance plays a crucial role in balancing asymmetrical objects when they received mechanistic explanations combined with either primary or secondary anomalous evidence.

The current findings are consistent with previous research in two ways. First, young children develop theories about how to balance different types of objects through their informal daily experiences (Karmiloff-Smith and Inhelder, 1974; Siegler and Chen, 1998). Most children had an intuitive theory about how to balance symmetrical and asymmetrical objects prior to the intervention, with only 25% (26 out of 102) of children categorized as having no theory. Second, a greater number of children from the experimental conditions advanced to a mass theory after the intervention compared to the control condition, particularly when they possessed a theory, either a traditional or transition theory. Karmiloff-Smith and Inhelder (1974), showed that the presence of a theory aided children

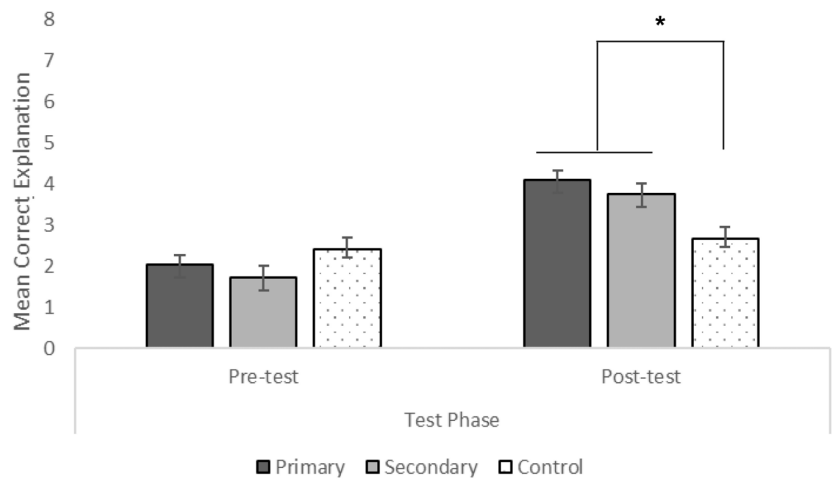


FIGURE 5 | Mean explanation responses out of 8 as a function of test phase and condition (*$p < 0.001$).

in exploring how to balance asymmetrical beams. Similarly, when given the chance to balance objects over several days, possessing a theory prior to beginning the task promoted the advancement through Siegler's (1976) four rules about balance relations (Siegler and Chen, 1998). The authors postulated that the 4- and 5-year-olds who had a theory appreciated the value of systematicity and were more likely to form and follow more advanced rules than children who did not have a theory.

The current research unifies previous research on children's balance theories. Previous research showed that, with time or explicit instructions, 5-year-olds can consider the effect of distance when balancing asymmetrical blocks (Siegler and Chen, 1998; Pine et al., 1999) whilst other studies found that when 5-year-olds witnessed evidence without an explanation they were not able to learn how to balance asymmetrical blocks (Bonawitz et al., 2012; Ganea et al., 2017). Our finding of an intermediate theoretical phase, Transition Theorists, is important in resolving these divergent findings. Specifically, we found that at post-test, 50% of children were categorized as Transition Theorists, demonstrating the prominence of the Transition phase stage. Further, many children who were Traditional Theorists or did not have a theory at pre-test were categorized as Transition Theorists at post-test. As such, the inclusion of this phase better captures the progression of 5-years-old ability to form theories on how to balance unevenly weighted blocks. We found significant learning for this younger age group, with similar degrees of learning from both primary and secondary sources of evidence, which indicates that 5-year-olds can develop theories about balance that incorporate distance when tested immediately following explicit instruction.

These findings also showed that credible sources of evidence that provide anomalous examples combined with explanation benefit belief revision. Our experimental conditions were structured so that each participant received the same number of examples and the connections between the examples and the explanation were matched. Multiple exemplars of anomalous evidence can promote deeper understanding of the target concept (Gentner, 1983, 1989) by reducing invalid inferences (Vosniadou and Skopeliti, 2019) as well as aiding in visualizing the explanation (Vosniadou, 1989; Brown, 1992, 1993). Further, the explanations not only highlighted the significance of a novel variable (i.e., distance), which is challenging for younger children to consider on their own, but it also highlighted its causal role (Siegler and Chen, 1998). This work supports Brown's (1992) conclusion that examples must be clearly analogous to the concept in question and that they be presented in a connected sequence referencing an explicit mechanistic model. In other words, the structure of our experimental conditions, with clearly analogous anomalous evidence situated within a causal explanatory framework, was the probable mechanism that facilitated the creation and revision of theories for younger children.

While the current work shows that children learn regardless of the source of evidence, this domain is not one with longstanding prior beliefs. That is, by early elementary school most children have progressed from a center theorist to a mass theorist (Karmiloff-Smith and Inhelder, 1974). This can be contrasted with other domains where naive beliefs persist into adulthood, such as the belief that heavier objects fall faster than light objects (Kavanagh and Sneider, 2007). Children's belief revision is likely influenced by the interaction between the strength of their prior beliefs and the complexity of the evidence presented. Children may have markedly diverse responses to different sources of evidence across various knowledge domains. An open question remains whether children would benefit from multiple exemplars alone, or if the explanations provided an additive benefit, especially when children have strong competing prior beliefs. Further research is needed to explore children's responses to different sources of evidence about various scientific concepts.

Additionally, our study was a single intervention completed in a one-on-one setting. Proceeding from belief revision to conceptual change is a gradual process (Vosniadou, 2002) as new beliefs do not simply replace old beliefs. Prior conceptions continue to coexist with newly learned information where either may be utilized depending on the situation (Shtulman and Valcarcel, 2012; Shtulman and Lombrozo, 2016). Given the study design we cannot discern if children maintain their learning at a delay. Therefore, it is uncertain whether children underwent conceptual change or short-term belief revision. More research is needed to examine these results over a longer period of time. Future work should also look at how these questions transfer to a naturalistic setting, to determine the implications for early science education.

CONCLUSION

This research adds to our understanding of how children develop scientific theories and the incremental changes that these theories undergo as children are exposed to anomalous evidence and causal explanations. With increased exposure to reliable sources of evidence and accompanying explanations, children can reach an understanding of scientific principles, at least in the domain of balance, earlier than previously thought (Karmiloff-Smith and Inhelder, 1974). Such findings justify the call for science education in the early years (Bowman et al., 2001; Gelman and Brenneman, 2004; Eshach and Fried, 2005; Duschl et al., 2007; Morgan et al., 2016). Promoting an earlier understanding of these concepts will serve as a foundation for more advanced concepts as children progress through school (Hardy et al., 2006). Indeed, scientific knowledge at preschool predicts children's science achievement in later grades (Morgan et al., 2016). Further research on how children's scientific understanding develops and what approaches improve their understanding can help conceptualize how to build effective early science education.

DATA AVAILABILITY STATEMENT

The data available on the Open Science Framework website: https://osf.io/84xap/?view_only=6747d7eec0054c5aa22951a8a0b95101.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Toronto Ethics Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

PG: conceptualization, methodology, writing – review and editing, supervision, and funding acquisition. NL and VV: conceptualization, methodology, investigation, analysis, and writing – original draft. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank the children and families who participated in this research and the Language and Learning Lab members at the University of Toronto for their help with recruitment and coding. This research was supported by an Insight grant from the Social Sciences and Humanities Research Council of Canada (SSHRC) to PG.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01503/full#supplementary-material

REFERENCES

Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edn. Hoboken, NJ: Wiley.

Baillargeon, R. (2002). "The acquisition of physical knowledge in infancy: a summary in eight lessons," in *Blackwell Handbook of Childhood Cognitive Development*, ed. U. Goswami (Malden, MA: Blackwell Publishers Ltd), 46–83.

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., and Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cogn. Psychol.* 64, 215–234. doi: 10.1016/j.cogpsych.2011.12.002

Bowman, B., Donovan, M., and Burns, M. (2001). *Eager to Learn: Educating our Preschoolers*. Washington, DC: National Academy Press.

Brewer, W. F., Chinn, C. A., and Samarapungavan, A. (1998). Explanation in scientists and children. *Minds Mach.* 8, 119–136. doi: 10.1023/A:1008242619231

Brown, D. E. (1992). Using examples and analogies to remediate misconceptions in physics: factors influencing conceptual change. *J. Res. Sci. Teach.* 29, 17–34. doi: 10.1002/tea.3660290104

Brown, D. E. (1993). Refocusing core intuitions: a concretizing role for analogy in conceptual change. *J. Res. Sci. Teach.* 30, 1273–1290. doi: 10.1002/tea.3660301009

Chinn, C. A., and Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: a theoretical framework and implications for science instruction. *Rev. Educ. Res.* 63, 1–49. doi: 10.3102/00346543063001001

Clement, J. J. (2013). "Roles for explanatory models and analogies in conceptual change," in *International Handbook of Research on Conceptual Change*, ed. S. Vosniadou (Abingdon: Routledge), 412–446.

Dunbar, K., and Klahr, D. (1989). "Developmental differences in scientific discovery processes," in *Complex Information Processing: The Impact of Herbert A. Simon*, eds D. Klahr, and K. Kotovsky (Lawrence: Erlbaum Associates, Inc), 109–143.

Duschl, R. A., Richard, A., Schweingruber, H. A., Shouse, A. W., National Research Council (U.S.), Committee on Science Learning, Kindergarten Through Eighth Grade, et al. (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8*. Washington, D.C: National Academies Press.

Eshach, H., and Fried, M. N. (2005). Should science be taught in early childhood? *J. Sci. Educ. Technol.* 14, 315–336. doi: 10.1007/s10956-005-7198-9

Ganea, P. A., Larsen, N., and Venkadasalam, V. (2020). The role of alternative theories and anomalous evidence in children's scientific belief revision. *Child Development*.

Ganea, P. A., Simons-Ghafari, K., and Walker, C. M. (2017). Weighing the evidence: promoting belief revision through storybooks. *Paper Symposium Presented at the Society for Research in Child Development*, Austin, TX.

Gelman, R., and Brenneman, K. (2004). Science learning pathways for young children. *Early Childhood Res. Q.* 19, 150–158. doi: 10.1016/j.ecresq.2004.01.009

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. *Cogn. Sci.* 7, 155–170. doi: 10.1207/s15516709cog0702_3

Gentner, D. (1989). "Analogical learning," in *Similarity and Analogical Reasoning*, eds S. Vosniadou, and A. Ortony (Cambridge: Cambridge University Press), 199–241. doi: 10.1017/cbo9780511529863.011

Hardy, I., Jonen, A., Möller, K., and Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking. *J. Educ. Psychol.* 98, 307–326. doi: 10.1037/0022-0663.98.2.307

Harris, P. L., and Corriveau, K. H. (2014). "Learning from testimony about religion and science," in *Current Issues in Developmental Psychology. Trust and Skepticism: Children's Selective Learning From Testimony*, eds E. J. Robinson, and S. Einav (Hove: Psychology Press), 28–41.

Harris, P. L., and Koenig, M. A. (2006). Trust in testimony: how children learn about science and religion. *Child Dev.* 77, 505–524. doi: 10.1111/j.1467-8624.2006.00886.x

Howell, D. (2007). *Statistical Methods for Psychology*, 6 Edn. Belmont, CA: Wadsworth.

Karmiloff-Smith, A., and Inhelder, B. (1974). If you want a theory, get a theory. *Cognition* 3, 195–212. doi: 10.1016/0010-0277(74)90008-0

Kavanagh, C., and Sneider, C. (2007). Learning about gravity: I. Free fall: a guide for teachers and curriculum developers. *Astron. Educ. Rev.* 5, 21–52. doi: 10.3847/AER2006018

Kelemen, D. (2019). The magic of mechanism: explanation-based instruction on counterintuitive concepts in early childhood. *Persp. Psychol. Sci.* 14, 510–522. doi: 10.1177/1745691619827011

Klahr, D., and Dunbar, K. (1988). Dual space search during scientific reasoning. *Cogn. Sci.* 12, 1–48. doi: 10.1207/s15516709cog1201_1

Koerber, S., Osterhaus, C., and Sodian, B. (2017). Diagrams support revision of prior belief in primary-school children. *Frontline Learning Research* 5, 76–84. doi: 10.14786/flr.v5i1.265

Koslowski, B. (1996). *Theory and Evidence: The Development of Scientific Reasoning*. Cambridge, MA: The MIT Press.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychol. Rev.* 96, 674–689. doi: 10.1037/0033-295X.96.4.674

Kuhn, D., Amsel, E., and O'Loughlin, M. (1988). *The Development of Scientific Thinking Skills*. New York, NY: Academic Press.

Morgan, P. L., Farkas, G., Hillemeier, M. M., and Maczuga, S. (2016). Science achievement gaps begin very early, persist, and are largely explained by modifiable factors. *Educ. Res.* 45, 18–35. doi: 10.3102/0013189X16633182

National Institutes of Health, (2015). *NIH Toolbox Picture Vocabulary Test. Used with Permission*. Bethesda, MD: National Institutes of Health.

Pine, K. J., Messer, D. J., and Godfrey, K. (1999). The teachability of children with naive theories: an exploration of the effects of two teaching methods. *Br. J. Educ. Psychol.* 69, 201–211. doi: 10.1348/000709999157671

Renken, M. D., and Nunez, N. (2010). Evidence for improved conclusion accuracy after reading about rather than conducting a belief—inconsistent simple physics experiment. *Appl. Cogn. Psychol.* 24, 792–811. doi: 10.1002/acp.1587

Renken, M. D., and Nunez, N. (2013). Computer simulations and clear observations do not guarantee conceptual understanding. *Learn. Instruct.* 23, 10–23. doi: 10.1016/j.learninstruc.2012.08.006

Sandoval, W. A., and Çam, A. (2011). Elementary children's judgments of the epistemic status of sources of justification. *Sci. Educ.* 95, 383–408. doi: 10.1002/sce.20426

Sandoval, W. A., Sodian, B., Koerber, S., and Wong, J. (2014). Developing children's early competencies to engage with science. *Educ. Psychol.* 49, 139–152. doi: 10.1080/00461520.2014.917589

Schauble, L. (1990). Belief revision in children: the role of prior knowledge and strategies for generating evidence. *J. Exp. Child Psychol.* 49, 31–57. doi: 10.1016/0022-0965(90)90048-D

Shtulman, A. (2017). *Scienceblind: Why Our Intuitive Theories About the World Are So Often Wrong*. New York, NY: Basic Books.

Shtulman, A., and Lombrozo, T. (2016). "Bundles of contradiction: a coexistence view of conceptual change," in *Core Knowledge and Conceptual Change*, eds D. Barner, and A. Baron (Oxford: Oxford University Press), 49–67.

Shtulman, A., and Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition* 124, 209–215. doi: 10.1016/j.cognition.2012.04.005

Siegler, R. S. (1976). Three aspects of cognitive development. *Cogn. Psychol.* 8, 481–520. doi: 10.1016/0010-0285(76)90016-5

Siegler, R. S., and Chen, Z. (1998). Developmental differences in rule learning: a microgenetic analysis. *Cogn. Psychol.* 36, 273–310. doi: 10.1006/cogp.1998.0686

Tabachnick, B. G., and Fidell, L. S. (2012). *Using Multivariate Statistics*. Boston, MA: Pearson.

Tolmie, A. K., Ghazali, Z., and Morris, S. (2016). Children's science learning: a core skills approach. *Br. J. Educ. Psychol.* 86, 481–497. doi: 10.1111/bjep.12119

Vokey, J. R. (2003). Multiway frequency analysis for experimental psychologists. *Can. J. Exp. Psychol.* 57, 257–264. doi: 10.1037/h0087429

Vosniadou, S. (1989). "Analogical reasoning as a mechanism in knowledge acquisition: a developmental perspective," in *Similarity and Analogical Reasoning*, eds S. Vosniadou, and A. Ortony (Cambridge: Cambridge University Press), 413–437. doi: 10.1017/cbo9780511529863.020

Vosniadou, S. (2002). "On the nature of naive physics," in *Reconsidering Conceptual Change: Issues in Theory and Practice*, eds M. Limon, and L. Mason (Dordrecht: Springer), 61–76. doi: 10.1007/0-306-47637-1_3

Vosniadou, S., and Skopeliti, I. (2019). Evaluating the effects of analogy enriched text on the learning of science: the importance of learning indexes. *J. Res. Sci. Teach.* 56, 732–764. doi: 10.1002/tea.21523

Wickens, T. D. (1989). *Multiway Contingency Tables Analysis for the Social Sciences*. Hillsdale, NJ: Erlbaum.

Zimmerman, C., and Klahr, D. (2018). Development of scientific thinking. *Stev. Handbook Exp. Psychol. Cogn. Neurosci.* 4, 1–25.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Larsen, Venkadasalam and Ganea. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Psychology

ORIGINAL RESEARCH
published: 01 December 2020
doi: 10.3389/fpsyg.2020.531565

The Predictive Value of Children's Understanding of Indeterminacy and Confounding for Later Mastery of the Control-of-Variables Strategy

Sonja Peteranderl and Peter A. Edelsbrunner**

ETH Zurich, Zurich, Switzerland

Prior research has identified age 9–11 as a critical period for the development of the control-of-variables strategy (CVS). We examine the stability of interindividual differences in children's CVS skills with regard to their precursor skills during this critical developmental period. To this end, we relate two precursor skills of CVS at age 9 to four skills constituting fully developed CVS more than 2 years later, controlling for children's more general cognitive development. Note that $N = 170$ second- to fourth-graders worked on multiple choice-assessments of their understanding of indeterminacy of evidence and of confounding. We find relations between these two precursor skills and children's CVS skills 2 years later at age 11 in *planning*, *identifying*, and *interpreting* controlled experiments, and in recognizing the inconclusiveness of confounded comparisons (*understanding*). In accordance with the perspective that both indeterminacy and confounding constitute critical, related yet distinct elements of CVS, both precursor skills contribute to the prediction of later CVS. Together, the two precursor skills can explain 39% of students' later CVS mastery. Overall, the understanding of indeterminacy is a stronger predictor of fully developed CVS than that of confounding. The understanding of confounding, however, is a better predictor of the more difficult CVS sub-skills of understanding the inconclusiveness of confounded comparisons, and of planning a correctly controlled experiment. Importantly, both precursor skills maintain interactive predictive strength when controlling for children's general cognitive abilities and reading comprehension, showing that the developmental dynamics of CVS and its precursor skills cannot be fully ascribed to general cognitive development. We discuss implications of these findings for theories about the development of CVS and broader scientific reasoning.

Keywords: control-of-variables-strategy, cognitive development, indeterminacy, confounding, longitudinal study, additive mixed models

OPEN ACCESS

Edited by:
Amanda C. Brandone, Lehigh University, United States

Reviewed by:
David Klahr, Carnegie Mellon University, United States
Susanne Koerber, University of Education Freiburg, Germany

****Correspondence:***
Sonja Peteranderl
sonja.peteranderl@ifv.ethz.gess.ch
Peter A. Edelsbrunner
peter.edelsbrunner@ifv.gess.ethz.ch

Specialty section:
This article was submitted to Developmental Psychology, a section of the journal Frontiers in Psychology

Received: *31 January 2020*
Accepted: *08 September 2020*
Published: *01 December 2020*

Citation:
Peteranderl S and Edelsbrunner PA (2020) The Predictive Value of Children's Understanding of Indeterminacy and Confounding for Later Mastery of the Control-of-Variables Strategy. Front. Psychol. 11:531565. doi: 10.3389/fpsyg.2020.531565

1. INTRODUCTION

Scientific reasoning, which is typically described as a cyclic process of intentional knowledge-seeking through an empirical research process, encompasses skills such as generating and testing hypotheses, conducting controlled experiments, and the data-based evaluation of these experiments (Klahr, 2000; Kuhn et al., 2000; Wilhelm and Beishuizen, 2003; Zimmerman, 2007).

Regarding experimentation, one crucial component is the control-of-variables strategy (CVS; Chen and Klahr, 1999). The CVS describes the technique to hold the levels of all variables constant except for the variable being investigated (Tschirgi, 1980; Chen and Klahr, 1999; Klahr, 2000; Dewey, 2002).

Although early developmental research indicated that children cannot develop understanding of CVS before early adolescence (Siegler et al., 1973; Tschirgi, 1980), later research indicated that precursor skills emerge already during childhood (Sodian et al., 1991; Bullock et al., 2009; Piekny and Maehler, 2013; Koerber and Osterhaus, 2019). We define precursor skills as the first emerging skills that build the foundation of more advanced and fully developed CVS. In this study, we examine whether and to what extent such precursor skills, in the present case children's understanding of indeterminacy of evidence and of confounding, can predict their mastery of more fully developed CVS skills 2 years later, and whether they have predictive value beyond children's general cognitive development.

1.1. Crucial Development Before Adolescence

Traditionally, educational and developmental researchers considered the experimentation skills of elementary school children to be deficient (Inhelder and Piaget, 1958; Klahr et al., 1993). The development of these skills was said to not emerge before adolescence (Inhelder and Piaget, 1958; Klahr et al., 1993). By contrast, recent research has revealed increasing evidence for a crucial period of development of the understanding of the CVS before adolenscence (e.g., Sodian et al., 1991; Bullock and Ziegler, 1999; Zimmerman, 2007; Bullock et al., 2009). Sodian et al. (1991) showed that already elementary school children could differentiate between controlled and confounded experiments. Bullock and Ziegler (1999) and Bullock et al. (2009) showed that children by the age of 8 years prefer controlled experiments over confounded experiments, but a spontaneous application of CVS has not been found (see also Zimmerman, 2007).

Specifically, Bullock and Ziegler (1999) delivered consistent findings on a task requiring the production and recognition of adequate tests for hypotheses and the relations between variables in the LOGIC study (Munich Longitudinal Study of the Genesis of Individual Competencies) with third to sixth graders. The children were asked to select from different comparisons that were either confounded or controlled to test a given hypothesis. By the age of ∼8 years old, children preferred the conclusive comparisons over the confounded comparisons. In addition, more than 50% of the fourth graders, ∼80% of the fifth graders and almost all of the sixth graders justified their choices by referring to the control of variables. Although children focus primarily on reasonable hypotheses (Klahr et al., 1993), they often recognize good comparisons if they are coherent with their initial beliefs (Sodian et al., 1991; Gopnik and Schulz, 2004; Croker and Buchanan, 2011), or they generate hypotheses to fulfill their prior theories (Kuhn et al., 1995; Schauble, 1996; Croker and Buchanan, 2011). These findings indicate some understanding of the CVS already in childhood, with increased development between 8 and 12 years of age.

This does however not imply that all individuals develop perfect CVS understanding before adolescence. Various studies (e.g., Kuhn, 2007) have shown that in some individuals, understanding of the CVS does not develop up to adulthood. Bullock et al. (2009) found that adults struggled with a metaconceptual understanding of alternative theories within experimental design (for review, see also Zimmerman and Croker, 2013; Zimmerman and Klahr, 2018).

Based on these results, we describe age 9–11 as a crucial developmental period. By crucial development period, we mean that increased development takes place during this period, and that there are shifts in the kinds of tasks that children are able to master before and after this period. The described evidence indicates that even some adults do not fully grasp the CVS. Hence, we presume that for some individuals who do not develop CVS during this period, further development might also be limited in the years after.

1.2. Sources of Development of CVS and Broader Experimentation Skills

There are at least three different sources that can contribute to children's development of CVS and broader experimentation skills. Even though not all children develop CVS on their own (Zimmerman, 2007), its development can be supported by implicit and explicit educational interventions (Chen and Klahr, 1999; Schalk et al., 2019). Schwichow et al. (2016b) summarized the findings from 72 CVS intervention studies and found a mean overall effect size of $g = 0.61$, with some studies indicating that even 6-year-olds can benefit from training.

Besides direct training, a more general factor that contributes to children's development of CVS is their general cognitive development. CVS and related experimentation skills do not develop fully independently of other cognitive abilities (for an overview, see Edelsbrunner et al., in Press). For example, CVS and related skills have been found to be associated with general reasoning skills (Mayer et al., 2014; Wagensveld et al., 2015), and with verbal skills and reading comprehension (Siler et al., 2010; Mayer et al., 2014; van der Graaf et al., 2016; Osterhaus et al., 2017).

Finally, various skills related to CVS and broader scientific reasoning might represent sources of development for each other through mutual developmental relations. CVS and further aspects of scientific reasoning themselves are usually intercorrelated (e.g., Mayer et al., 2014). Such interrelations relate to the question of whether scientific reasoning should be described as a unidimensional construct, or as a construct that incorporates multiple dimensions that operate and develop in parallel (Zimmerman, 2007). Koerber et al. (2015) proposed a conceptual model based on the idea that the core of scientific reasoning is the ability to differentiate and coordinate theories and evidence (Kuhn, 2010). Based on this idea, they predicted a unidimensional psychometric structure of scientific reasoning, describing evidence for this notion in Mayer et al. (2014) and Koerber et al. (2015), and Koerber and Osterhaus (2019) based

on finding adequate itemfit in Rasch analyses. Edelsbrunner and Dablander (2018), however, based on data simulations and psychometric considerations, argued that itemfit provides limited information for distinguishing a common core from further sources of intercorrelations between different skills related to scientific reasoning. Testing different assumptions about the cognitive abilities underlying a set of intercorrelated tasks is generally very difficult based on cross-sectional data (van Bork et al., 2019; VanderWeele and Batty, 2020). More fruitful information regarding the question of whether it is useful to model skills related to scientific reasoning unidimensionally or multidimensionally might be gathered from looking into the developmental interplay of multiple skills over time. For example, if multiple skills related to scientific reasoning show additive or interactive developmental and predictive patterns regarding later skills, this would imply that the former skills should be conceptualized and modeled separately. If, however, multiple skills related to scientific reasoning exhibit interchangeable developmental patterns or predictive value for later skills, treating these as a unitary construct would not imply loss of information.

1.3. Different Sub-skills of CVS

Different types of tasks have been used to assess CVS, with tasks that assess similar skills sometimes receiving different labels across studies. In the present study, we follow a distinction of CVS sub-skills that has been developed by Schwichow et al. (2016a) based on a definition of CVS used in a seminal study by Chen and Klahr (1999).

According to the distinction by Schwichow et al. (2016a), CVS incorporates four sub-skills: *interpreting* controlled experiments, *identifying* controlled experiments, *understanding* the indeterminacy of confounded comparisons, and *planning* controlled experiments. More specifically, *interpreting* describes the ability to interpret a controlled experiment based on the outcome. *Identifying* describes the ability to select a suitable comparison according to a specific hypothesis; *understanding* declares knowledge about the indeterminacy of confounded experiments; in other words, the knowledge that no valid conclusion can be drawn from a confounded experiment. *Planning* describes the capability to build a comparison according to a given initial hypothesis based on provided variables. A comparison between the labels that were given to tasks that involve similar skills helps elucidate the differences between the four sub-skills, and why these distinctions in our perspective are informative.

In a recent study, Koerber and Osterhaus (2019) labeled tasks in which children had to interpret controlled or confounded experiments *data interpretation*. In the distinction by Schwichow et al. (2016a) and our study, these tasks would fall under two different categories. When children have to draw the correct conclusion from a controlled experiment, we call the relevant CVS sub-skill *interpretation*. In the case in which children have to interpret a confounded comparison, however, they have to understand that no conclusive conclusion is possible because of the confounding. We call the relevant CVS sub-skill *understanding*. Schwichow et al. (2016a) found that the latter appears to be much more difficult for students. In a more recent study, Schwichow et al. (2020) found that understanding is the most difficult sub-skill and in analyses of subgroups of students, they found that it might premise mastering the other more procedural sub-skills. Given the differential difficulty and information that these two sub-skills seem to provide, we also distinguish between the controlled and confounded cases and the involved sub-skills.

The remaining two sub-skills, *identifying* and *planning*, can be distinguished by comparing them to tasks in the longitudinal study by Bullock et al. (2009). In tasks that Bullock et al. (2009) labeled choice-tasks, children had to select the correct, controlled comparison among different proposals for setting up an experiment. In the present study, we refer to this kind of task as assessing the precursor skill of confounding we will further explain this in the next section, and in our conceptualization of fully developed CVS it receives the label *identifying* . Hence we see this as a development of the precursor to a fully developed CVS sub-skill. In tasks that Bullock et al. (2009) labeled production-tasks, children had to actively set up experimental comparisons on their own. In the present study, such tasks receive the label *planning*. Bullock et al. (2009) and further studies (see Zimmerman, 2007) found that production tasks were more difficult and children's underlying skills developed later than those on choice-tasks. We therefore also distinguish between these two different tasks and the underlying sub-skills.

1.4. Precursor Skills of CVS

With regard to the kinds of tasks that children can solve, based on prior findings we distinguish between precursor skills of CVS that many children develop in earlier childhood, and fully developed CVS that typically does not develop before this period. Two precursor skills of more fully developed CVS are the understanding of indeterminacy and that of confounding.

Indeterminacy refers to the understanding of whether available evidence is sufficient to warrant a specific conclusion (Klahr and Chen, 2003). For example, when a toy consists of plugged round parts and the question is which of two boxes the parts were taken from, but both boxes contain round parts, then the evidence is indeterminate (Piéraut-Le Bonniec, 1980). In experimental design, the indeterminacy principle pertains to the focal variable about which the causal status is in question. In order to produce determinate evidence, the focal variable has to be varied such that it can yield determinate evidence. An example for indeterminate experimental evidence is when an object that smells strongly to humans is hidden to find out whether a German shorthaired pointer-dog can smell better than humans: Even humans could have smelled the object; therefore, if the dog finds it, this does not denote determinate evidence, and no conclusion can be drawn. Indeterminacy makes the first puzzle piece to later mastery of CVS: Before the control of confounding variables is considered, the right focal variable has to be varied in a manner such that conclusions about the question at hand will be possible. Piekny and Maehler (2013) investigated the understanding of indeterminacy by asking children to design an experiment, employing a task by Sodian et al. (1991). They asked 4- to 12-year-old children to choose between two mice

houses one with a big opening and one with a small opening in order to feed both (the small and the large) mouses (problem 1) and find out whether the mouse that went inside to eat was either big or small (problem 2). Their results showed that more than 50% of the 5-year old children could solve correctly problem 1 but failed in solving problem 2.

A second precursor skill is the understanding of confounding. This is the second step toward fully developed CVS: In addition to correctly varying the focal variable, confounding variables have to be controlled. As precursor skill, we refer here to recognition of situations in which confounding variables are correctly controlled, or not. Multiple studies have shown that such recognition can be triggered in many children in simple tasks, whereas active or spontaneous application of variable control is comparably rarely found in children before the outlined crucial developmental period before adolescence that we have explained in the introduction (Kuhn, 1989; Bullock and Ziegler, 1999; Bullock et al., 2009; van der Graaf et al., 2015; Koerber and Osterhaus, 2019).

Although the understanding of indeterminacy and confounding represent two early facets of CVS, mastery of CVS encompasses all four sub-skills. These require more active production and deep conceptual understanding of the role of CVS. In the present study, we examine the predictive value of these two precursor skills for mastery of CVS encompassing the four sub-skills 2 years later.

1.5. The Present Study

The present study examined to which degree children's understanding of *indeterminacy* and of *confounding* can predict their mastery of the CVS 2 years later, after most of the crucial developmental period for CVS. We also examined whether the developmental patterns taking place during this period can be discerned from children's more general cognitive development. Specifically, we asked (1) to what extent the understanding of indeterminacy and confounding can predict later mastery of CVS, (2) whether the predictive strength of the precursor skills remains robust when taking into account children's general cognitive abilities and reading comprehension as covariates, (3) whether and how the precursor skills' predictive strength differs between the four CVS sub-skills of *planning*, *identifying*, and *interpreting* controlled experiments, and recognizing the inconclusiveness of confounded experiments (*understanding*), and again checking (4) whether the predictive relations hold when taking into account the covariates.

To this end, we assessed a sample of primary school children from Switzerland twice. The sample was part of the Swiss MINT Study, a large-scale longitudinal study on the effects of early Physics education conducted at ETH Zurich (for more details see https://educ.ethz.ch/lernzentren/mint-lernzentrum/Schweizer_MINT_Studie.html). The tasks applied in the present study to assess CVS and its sub-skills stem from Peteranderl (2019), who developed and evaluated a new inventory for measuring experimentation skills with a focus on fifth and sixth graders. This inventory covers the assessment of all four sub-skills of CVS, as summarized by Schwichow et al. (2016a), among the assessment of other skills of scientific reasoning, such as scientific argumentation or failure in experimentation because of incorrect preconceptions about experimentation, as summarized by Schauble et al. (1991) and Siler and Klahr (2012).

In the present study, the children were in second to fourth grade and on average 9 years old at the first assessment, meaning that this assessment took place in the beginning of the crucial developmental period 9–11. The second assessment took place 2 years later, when children were in fifth to sixth grade, about 11 years old on average, such that this assessment took place toward the end of the crucial developmental period.

At the first assessment, we assess children's understanding of indeterminacy and confounding, and at the second assessment the four CVS sub-skills, as well as general reasoning and reading comprehension. With this study design, we aimed at estimating the stability between the precursor skills and later CVS. Do children remain relatively stable from the precursor skills to later mastery of CVS, or do new substantial individual differences emerge during the crucial developmental period? In addition, we examine general cognitive abilities and reading comprehension, in order to test whether the predictive value of the two precursor skills represents specific dynamics that are distinct from children's more general cognitive development.

2. MATERIALS AND METHODS

2.1. Sample

The sample stemmed from two separate measurement points with a total sample size of 170 primary school children (90 female, 80 male) from 12 school classes in the German-speaking part of Switzerland. The children were in second ($n = 48$), third ($n = 54$), or fourth ($n = 68$) grade at the first assessment ($M_{age} = 9.07$, $SD = 0.95$), and in fifth ($n = 133$) or sixth ($n = 37$) grade at the second data assessment ($M_{age} = 11.18$, $SD = 0.46$). There was a median time gap (we report medians for skewed data) of 23 months between the two data assessments, with a range of 10–34 months.

2.2. Assessment of Precursor Skills

The assessments took place within the framework of the Swiss MINT Study, a longitudinal study investigating the impact of early science education on children's later academic and cognitive development (Edelsbrunner et al., 2015, 2018; Schalk et al., 2019).

At the first assessment (when children were in the second, third, or fourth grade; for an example, see 1), the children were assessed with the precursor skills assessment. This tool was a questionnaire encompassing 10 multiple-choice items (internal consistency reliability estimate: McDonald's omega = 0.81; for an explanation of omega; see Dunn et al., 2014) based on typical scenarios from the CVS literature (e.g., the Airplane- and Ramp-tasks; Bullock and Ziegler, 1999; Chen and Klahr, 1999) and similar scenarios.

The understanding of *indeterminacy* as one of two precursor skills was assessed by four items. Each of these items presented the children with a short story including the research question (e.g., "Which giraffe ate the carrot?") and each of the outcomes under different manipulations of the focal variable (e.g., "the big giraffe with her long neck can reach the tips of all trees"). In a

second step, the children were asked to select the correct variable in order to answer the initial research question by choosing the correct response within a multiple-choice response format. Children's mean score of correct answers on these items was used for analyses.

Confounding as the second precursor skill was assessed by six items. In these items, children were faced with a short story including different variables containing two levels each. In order to investigate a given research question referring to exactly one out of these variables (e.g., "what should Mr. Miller do to find our whether the form of the nose is important for how much fuel the airplane needs"), the children were then asked to select the correct response out of three possible multiple-choice responses. Children's mean score of correct answers on these items was used for analyses.

Figure 1 shows an example item of each precursor skill, indeterminacy (**Figure 1A**) and confounding (**Figure 1B**). The items have been validated and used in prior research on the development of CVS and the impact of educational interventions (Edelsbrunner et al., 2015, 2018; Schalk et al., 2019).

2.3. Assessment of the Control-of-Variables Strategy

At the second assessment (when children were in the fifth or sixth grade), the children were assessed with the CVS assessment. This tool stemmed from prior research investigating the impact of a CVS intervention in Swiss school children (Peteranderl, 2019). The assessment was a questionnaire encompassing 15 items with similar contexts as in the first tool (internal consistency reliability estimate of the overall test: omega = 0.84, of the sub-skill *interpreting*: omega = 0.91, of *identifying*: omega = 0.91, *understanding*: omega = 0.93, and *planning*: omega = 0.85). In this tool, the focus was on assessing all four sub-skills of CVS that can be distinguished based on the definition of Schwichow et al. (2016a) separately within different types of items. In **Figure 2**, items for the assessment of all four sub-skills of CVS are presented.

The CVS assessment encompassed four types of items for the assessment of the four sub-skills. However, children's solutions on the different item types contributed to scores of multiple of the four sub-skills. We first describe the four item-types, and then how children's scores on the four sub-skills were composed based on these. In all items, the children are faced with a research question and a hypothesis describing three to four different variables with two levels each. In the first type of items, *planning* (**Figure 2A**), the children are then asked to select the correct levels of four given variables according to the initial hypothesis in order to plan a conclusive experiment. In the second item type, *understanding* items (**Figure 2B**), the children are faced with an initial hypothesis and a suitable, but confounded experiment two variables are varied (instead of one variable) out of three given variables. They are asked to detect the inconclusiveness of this experiment and to select the correct response option reflecting the indeterminacy of confounding ("The children cannot tell for sure..."). In the third and fourth types of items, the *interpretation* and *identifying* items (**Figure 2C**), the children are faced with a research question related to three variables of interest. They are presented with four different sketches of four attempts to set up an experiment for examining the research question. The children are asked to select a suitable comparison to test the hypothesis (each item presents three hypotheses about the impact of the three variables on the outcome) reflecting the sub-skill *identifying*. The children are also asked to interpret the results of the outcome of the experiment with regard to the respective hypothesis, reflecting the sub-skill *interpreting*.

An overall CVS score of the second assessment was calculated based on all the items assessing the four CVS sub-skills, and four additional items. In the additional items, the children were asked for short written justifications about their reasoning in interpreting experiments. Ratings of these justifications together with children's scores on the four CVS sub-skills yielded an overall CVS score with a maximum of 35 points.

The scores for the four sub-skills were generated based on children's responses on the four described item types. The interpretation items contributed to the assessment of three sub-skills: *Interpreting* was constructed by drawing valid inferences from unconfounded comparisons in 10 cases. *Identifying* was constructed by the correct choice of an unconfounded comparison in the interpretation items in 10 cases. *Understanding* was constructed based on the detection of the indeterminacy of a confounded experiment in all of the interpretation items (in total 2 cases) and in all of the understanding items (in total 3 cases). The sub-skill *planning* was constructed by selecting a controlled experiment with regard to the initial hypothesis in the planning items (4 cases). The CVS assessment has been validated psychometrically, in cognitive interviews, and by asking students for additional open answers in order to probe construct validity (Peteranderl, 2019).

2.4. Covariates: Cognitive Abilities and Reading Comprehension

At the second assessment, we additionally measured children's general cognitive abilities and reading comprehension with standardized instruments. For measuring cognitive abilities, we used the numerical and figural analogies scales of the Germany-wide established cognitive abilities test for primary school children [Kognitiver Fähigkeitstest (KFT); Heller and Perleth, 2000]. In the numerical scale, students are presented with 20 items, each containing 4–5 numbers in a row following a specific rule. The students must ascertain this rule by recognizing the correct number continuing the row out of five provided answers (numbers). For this scale, 9 min was scheduled. In the figural scale, students are presented with 25 items containing pairs of figures that are related according to a specific rule that the student must determine. Afterwards, students must choose one of five provided answers (figures) that fits with another figure according to this rule. For this scale, 8 min was scheduled. Internal consistency reliability estimates were omega = 0.88 for the numerical scale and omega = 0.92 for the figural scale. The overall sum score from both scales was used as a covariate representing reasoning ability in the main analyses.

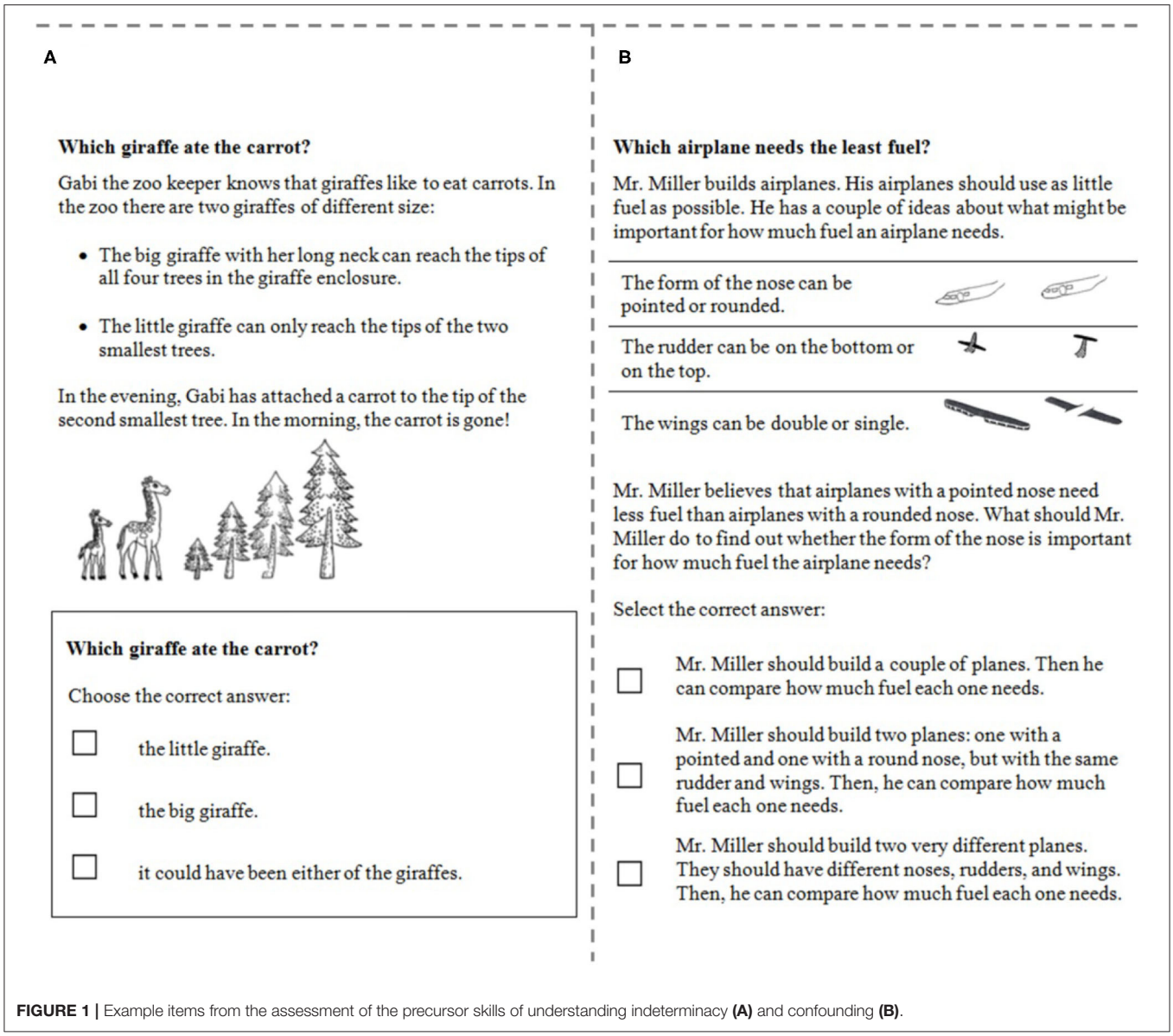


FIGURE 1 | Example items from the assessment of the precursor skills of understanding indeterminacy **(A)** and confounding **(B)**.

For measuring students' reading comprehension, we used the latest version of a standardized test instrument for fifth graders [Lesegeschwindigkeits-und-verstndnistest (LGVT, 2. Auflage); Schneider et al., 2017]. In this test, students have to read a text containing 2,161 words as far as they can within 6 min to measure reading speed. The text contains gaps with missing words that the students have to fill in by choosing the correct word from a selection of three words, measuring reading comprehension and reading accuracy. The internal consistency reliability estimate in this sample was omega = 0.80.

3. RESULTS

3.1. Descriptive Statistics

The descriptive statistics and intercorrelations of the main study variables are provided in **Tables 1**, **2**. The main study variables comprise the two precursor variables *indeterminacy*, understanding of conclusive testing, and *confounding*, understanding the control of nonfocal variables, at the first assessment. The main variables at the second assessment comprise the four sub-skills of CVS: *interpreting*, *identifying*, *understanding*, and *planning*, as well as overall *CVS* mastery and the two covariates.

In order to check that the distinctions between the two precursor skills at the first assessment, and between the four CVS sub-skills at the second assessment were psychometrically valid, we conducted two confirmatory factor analyses. In these analyses, we had the items load onto two (first assessment) or four (second assessment) latent variables representing the precursor skills and the four fully developed CVS sub-skills. The results of the analyses, depicted in **Figure 3**, indicated that the measurement models fit sufficiently well and that the two or four latent variables

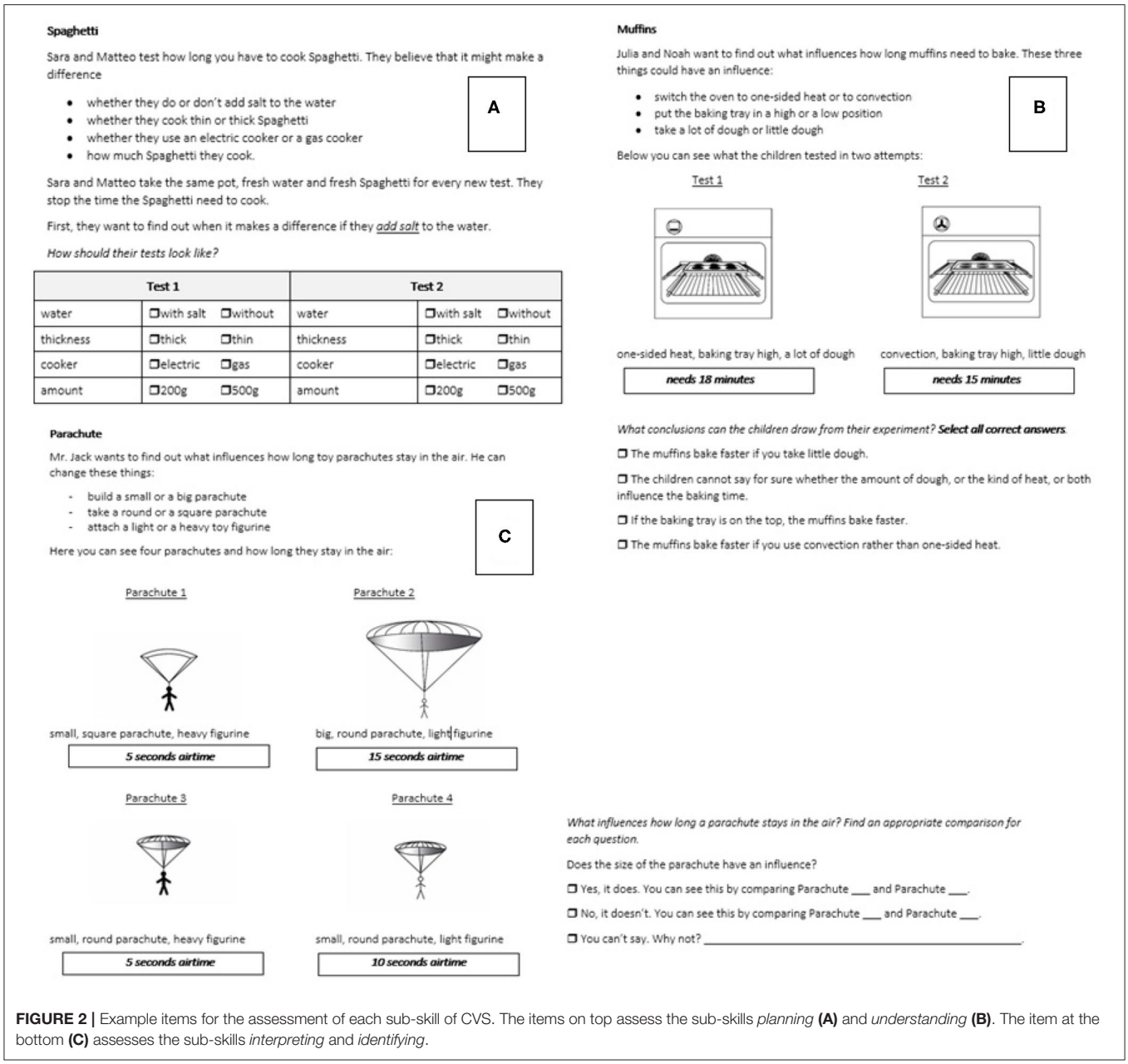


FIGURE 2 | Example items for the assessment of each sub-skill of CVS. The items on top assess the sub-skills *planning* **(A)** and *understanding* **(B)**. The item at the bottom **(C)** assesses the sub-skills *interpreting* and *identifying*.

showed moderate to substantial intercorrelations, which however were below 1.

3.2. The Predictive Strength of Precursor Skills for Later CVS

In order to examine the predictive strength of the two precursor skills for children's overall sum score on later CVS, we first investigated scatter plots (**Figure 4**). These indicated partially nonlinear relations between the two precursor skills and CVS. We therefore estimated additive mixed models, a regression technique that allows multilevel modeling, and particularly the capturing of nonlinear relations, whereas avoiding adjusting predictive terms unduly toward data outliers (Groll and Tutz, 2011; Wood, 2017). We fitted four models, the fit and explained variance of which are summarized in **Table 3**. We first estimated models with indeterminacy and confounding, respectively, as individual predictors for later CVS (**Table 3**: Models 1 and 2). We did not yet take into account classroom dependencies, because we were first of all interested in the predictive power of the two precursor skills when not yet deducing classroom differences that might overlap with the variances from these estimates. The two models indicated that indeterminacy alone could explain 29.7% of children's later CVS mastery ($F = 37.12$, $p < 0.001$), whereas confounding alone could explain 25% ($F = 11.83$, $p < 0.001$; for relative model fits, see **Table 3**).

TABLE 1 | Descriptive statistics of the main study variables.

	Maximum score	Mean (in %)	SD
First assessment			
Indeterminacy	1.00	0.54	0.30
Second grades ($n = 48$)		0.43	0.25
Third graders ($n = 54$)		0.60	0.31
Fourth graders ($n = 68$)		0.57	0.30
Confounding	1.00	0.37	0.27
Second grades ($n = 48$)		0.31	0.22
Third graders ($n = 54$)		0.45	0.31
Fourth graders ($n = 68$)		0.34	0.27
Second assessment			
CVS	35.00	13.58 (39%)	7.29
Identifying	10.00	4.33 (43%)	3.22
Interpreting	10.00	6.64 (66%)	2.81
Planning	4.00	0.98 (25%)	1.50
Understanding	5.00	0.54 (11%)	0.88
Cognitive abilities	45.00	30.44 (68%)	10.51
Reading comprehension	47.00	15.07 (32%)	10.06

Scores for separate school grades presented for indeterminacy and confounding. Scores for indeterminacy and confounding were scaled as mean scores ranging from 0 to 1, such that their mean scores can be interpreted as percentages, whereas scores on other measures indicate sum scores, with percentages in brackets.

TABLE 2 | Intercorrelations of the main study variables.

	Indet	Confo	CVS	Inter	Ident	Under	Plann	CogAb
Indeterminacy								
Confounding	0.47							
CVS	0.55	0.42						
Interpreting	0.34	0.17	0.73					
Identifying	0.46	0.30	0.85	0.54				
Understanding	0.28	0.33	0.49	0.22	0.20			
Planning	0.46	0.42	0.68	0.25	0.47	0.33		
CogAb	0.35	0.25	0.46	0.36	0.38	0.15	0.34	
ReadComp	0.37	0.27	0.52	0.34	0.44	0.20	0.42	0.48

Indet, indeterminacy; Confo, confounding; Inter, interpreting; Ident, identifying; Under, understanding; Plann, planning; CogAb, cognitive abilities; ReadComp, reading comprehension.

Combining both predictor variables and adding an interaction between the two variables (**Table 3**: Model 3) showed a small interaction ($F = 0.244$ $p = 0.035$) and main effects for both variables (indeterminacy: $F = 36.85$, $p < 0.001$, confounding: $F = 4.72$, $p = 0.001$), indicating that both variables add to the prediction of CVS beyond each other. In this combined model, both variables together managed to explain 39.2% of variation in later CVS mastery. We consider this model the most important one for our research question, because it shows the predictive power of the two precursor skills in combination. The relation of the two precursor skills with children's later CVS based on this model is depicted in **Figure 5**. A three-dimensional depiction is provided for visualizing the combined predictive value of both precursor skills. The figure can be read as follows: Children who are low on both predictor variables (dark area in lower front; both x- & y-axes close to 0) are estimated to solve about 20% of the later CVS items (z-axis). For children with low levels of indeterminacy (lower area of percentage solved on the axis labeled "Indeterminacy"), confounding has substantial positive predictive value: If children at least manage to solve many confounding-items, they are predicted to have relatively good estimated CVS-skill later on (about 50% solved items predicted; area high on confounding but low in indeterminacy in the right). The same holds vice versa: Children lower on the axis labeled "Confounding" but higher on the "Indeterminacy" axis have relatively good predicted CVS skill 2 years later, with about 55% solved items. The highest predicted later CVS skill, however, is reached by children who are high on both axes (area shaded in orange); for these children, estimates of later overall CVS skill are at about 80%. The interaction between the two predictor variables is for example visible for children who achieve moderate scores on confounding but low scores on indeterminacy; for these children, the estimated CVS skill 2 years later is higher than purely additive effects would indicate. Consequently, the surface within this area appears slightly elevated in comparison to the remaining surface patterns.

We added visual and inferential robustness checks to examine whether the grade in which children underwent the first assessment and the lag between the first and second assessment influenced these results. As visible from **Figure 6**, relations between the two precursor skills with later CVS were less strong if the first assessment took place already in second rather than third or fourth grade. Relations did not seem to differ, however, between those for whom the first assessment took place in third or fourth grade. Adding an interaction between the predictor terms of indeterminacy and confounding and the time lag between the first and second assessment did not show any interactions, both $p > 0.10$.

Finally, we examined the robustness of the predictive effects regarding children's covariates, as well as taking into account the classroom structure. We added a random intercept across teachers in order to control the intercept for the multilevel structure, and main effects of reasoning ability and reading comprehension (including their interactions with main effects; Simonsohn, 2019) to the model (**Table 3**: Model 4). In this model, controlling for classroom dependencies and the two covariates, the effects of indeterminacy and confounding became smaller, however the interaction effect, in particular, indicated that both variables still had predictive value beyond the variance shared with the classroom differences and covariates (indeterminacy: $F = 20.27$, $p < 0.001$; confounding: $F = 1.94$, $p = 0.100$; interaction: $F = 0.48$, $p = 0.014$).

Overall, these analyses indicate that both indeterminacy and confounding are substantial predictors of later CVS mastery and together they explain about 40% in the variance of later CVS (**Table 3**: Model 3). In addition, throughout the models, indeterminacy appeared as an overall stronger predictor of later CVS than confounding; however, both variables had predictive value beyond each other that remained intact when taking into

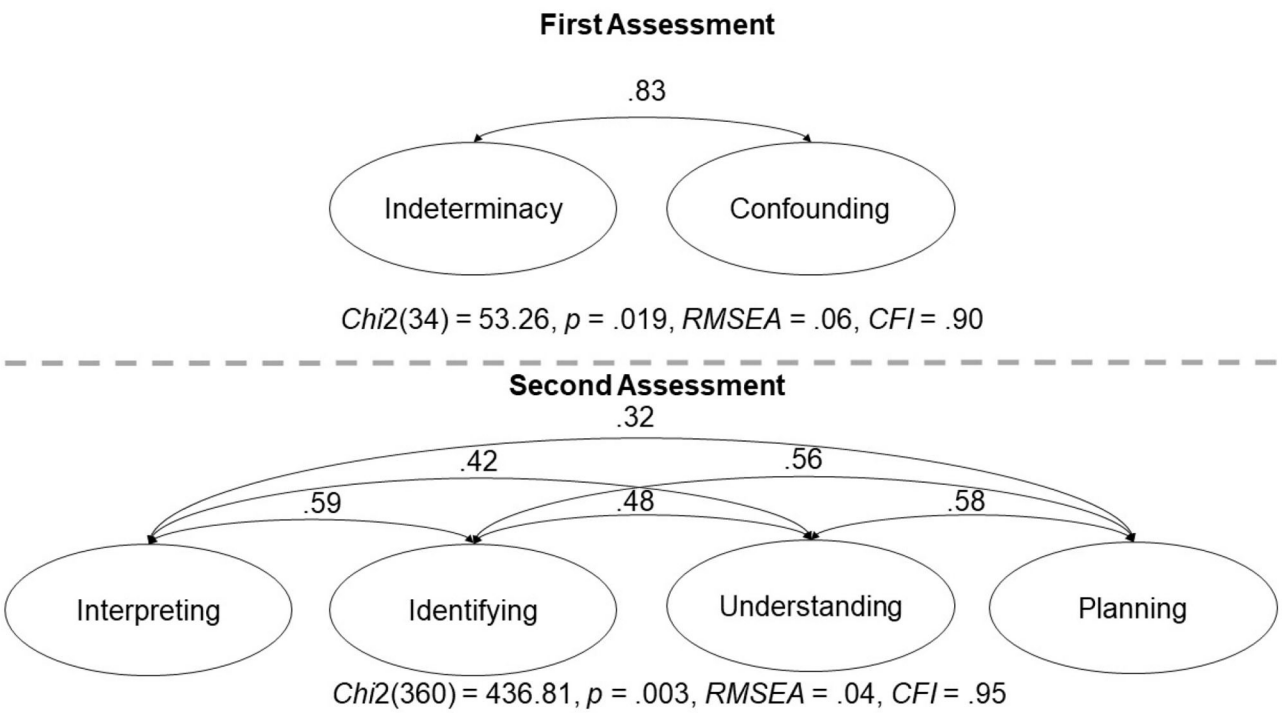


FIGURE 3 | Results from confirmatory factor analyses for the skills at the first (upper part) and second assessments (lower part), respectively. Indicator variables (individual items) omitted for visual clarity. All intercorrelations were significant at $p < 0.001$. When fixing the intercorrelation between inderterminacy and confounding to 1, the CFI deteriorated slightly to 0.89, and fixing any of the intercorrelations between the four CVS sub-skills from the second assessment to 1 substantially worsened model fits.

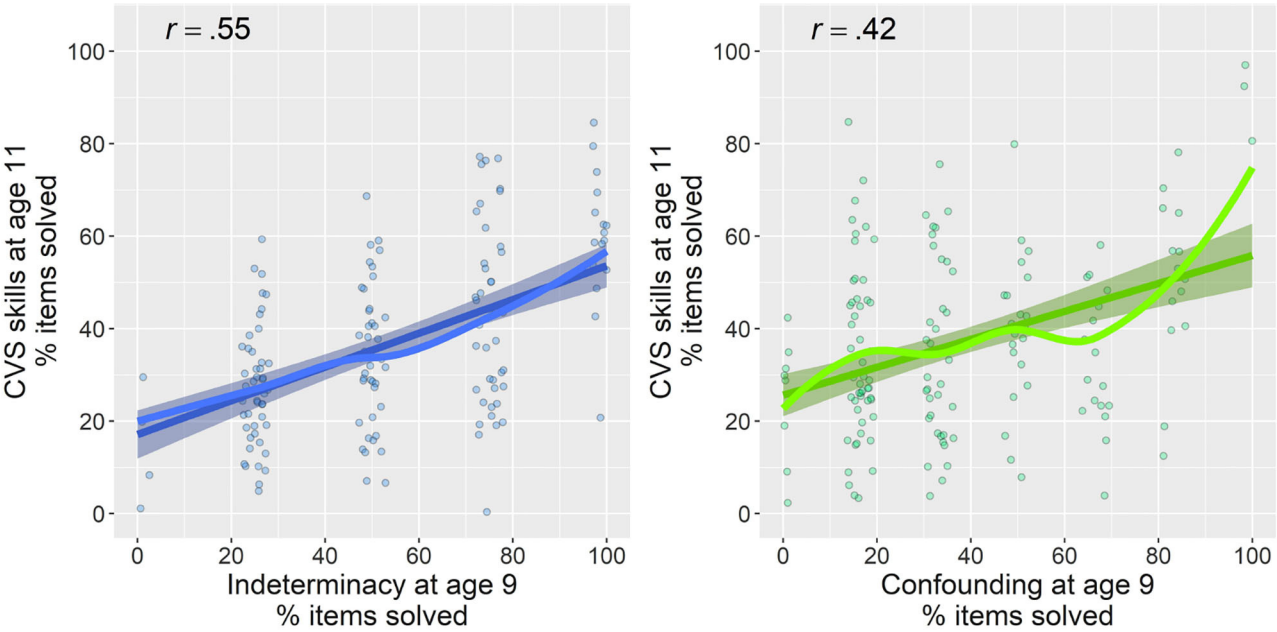


FIGURE 4 | Relations of indeterminacy **(left)** and confounding **(right)** at average age 9 with later overall CVS skill at average age 11. Percentage of items solved are plotted for all measures. Lines with 95% confidence band indicate linear trend, and lines without confidence bands indicate nonlinear (smooth) associations. Points are jittered for a better readability of scatterplots. Pearson correlations between variables are displayed in upper-left of each plot.

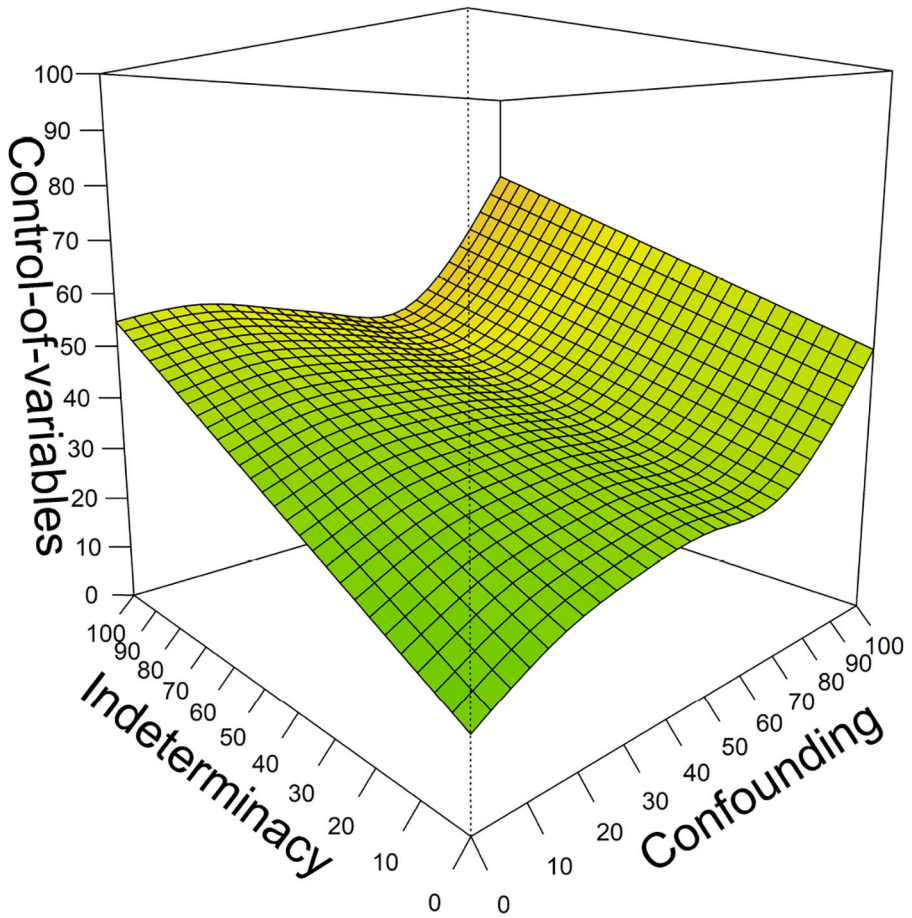


FIGURE 5 | The prediction of children's mastery of CVS from their understanding of indeterminacy and confounding 2 years earlier. All axes indicate model estimates of percentage of solved items on the respective measure. Areas in darker green indicate lower predicted later CVS skill, areas in brighter green moderate, and in orange highest later CVS skill.

TABLE 3 | Relative model fit and explained deviance for the four fitted general additive (mixed) models.

Model no.	Model description	AIC	BIC	% deviance explained
Model 1	Indeterminacy	1,460	1,472	29.7
Model 2	Confounding	1,474	1,495	24.9
Model 3	Indeterminacy × confounding	1,441	1,469	39.2
Model 4	Indeterminacy × confounding, covariates, teacher	1,393	1,505	61.6

Covariates were reasoning ability and reading comprehension, including their interactions with indeterminacy and confounding. teacher indicates a random intercept across teachers/classrooms.

account classroom dependencies and children's general cognitive development.

3.3. Variation in Predictive Strength Across Sub-skills

Associations between the precursor skills and the four later CVS sub-skills are depicted in **Figure 7**. All plots show the percentage of items solved for all four sub-skills of CVS. Lines with 95% confidence band indicate a linear trend, and lines without confidence band indicate the estimated nonlinear relation. The upper row shows the association between indeterminacy and the four sub-skills. The lower row shows the associations between confounding and the four sub-skills. According to these scatter plots, there are again some nonlinear relations between both precursor skills and the sub-skills of CVS. Hence, we fitted 16 additive mixed models. For each sub-skill, we estimated four models in line with the models described above. The results of the fitted models are summarized in **Table 4**. The eight models estimating both precursor skills as individual predictors (Models 1 and 2 for each sub-skill) indicated that indeterminacy alone

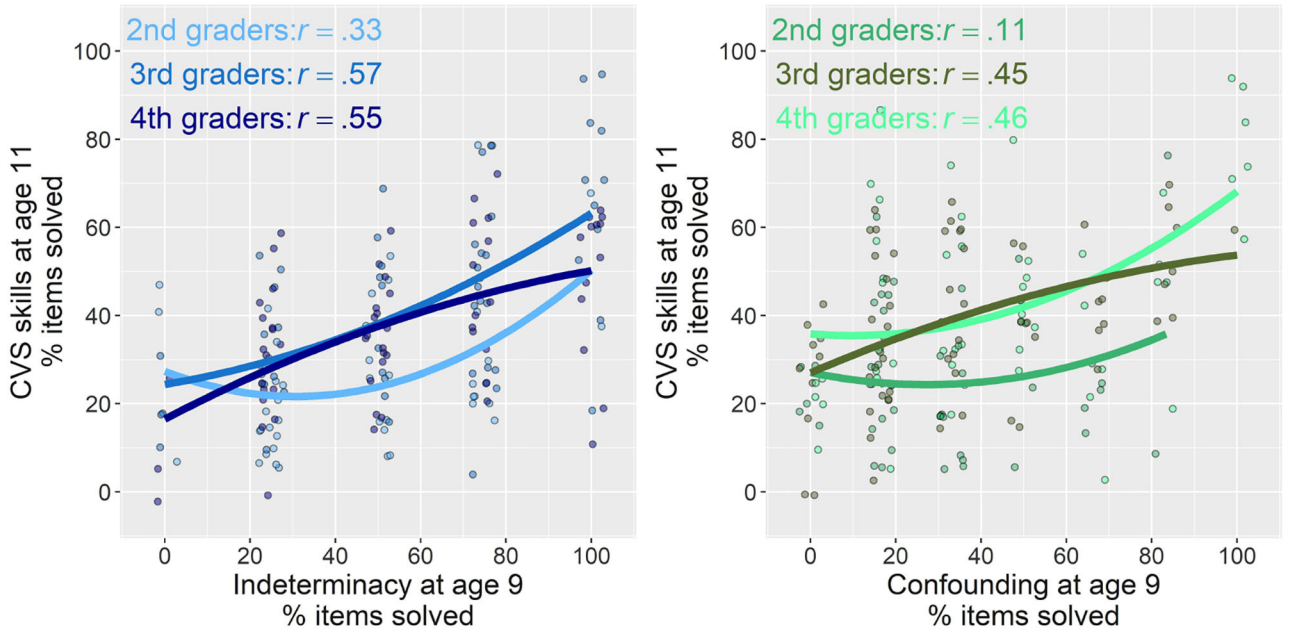


FIGURE 6 | Relations of indeterminacy **(left)** and confounding **(right)** at average age 9 with later overall CVS skill at average age 11 separate for the grades in which the first assessment took place. Pearson correlations between variables are displayed in upper left of each plot, whereas the regression lines indicate quadratic fit for a good balance of information and readability.

could explain between 9.3 and 22.3% of children's later variation in individual CVS sub-skills (all F's > 6.9, all p's < 0.001). The most deviance could be explained for the sub-skill *planning*. Confounding could explain between 5 and 24.5%, whereas the most explained deviance was estimated regarding the sub-skill *understanding* ($F = 14.91$, $p < 0.001$). This result was significant. The lowest explained deviance was estimated for *interpreting* ($F = 3.1$, $p = 0.28$). This result was not significant. Taking into account the interaction between the two predictors, we found for all four models (each Model no. 3 in **Table 4**) significant main effects for the precursor skill indeterminacy (all p's < 0.03). We found small significant interactions on the sub-skills *identifying, understanding, and planning* [identifying ($F = 0.30$, $p = 0.02$), understanding ($F = 1.95$, $p < 0.001$), planning ($F = 0.7$, $p = 0.006$)]. We did not find a significant interaction of the two predictors on interpreting ($F = 0.14$, $p = 0.1$).

Finally, we also tested the robustness of both precursor skills regarding the covariates, cognitive abilities, and reading comprehension. We fitted four models, taking into account the main effects of both covariates and additionally a random intercept across teachers in order to control for the multilevel structure. Results show no interactions between the two predictors, but robust main effects for indeterminacy regarding all four sub-skills (all p's < 0.02), indicating predictive value beyond shared variance with the covariates and the random intercept across teachers.

In sum, these results support the findings regarding mastery of overall CVS. Both precursor skill are substantial predictors for all four sub-skills of CVS. Indeterminacy shows stronger predictive value when taking into account children's covariates and classroom dependencies, and stronger predictive value for the sub-skill *interpreting*. Regarding the three other sub-skills, small interactions indicate that both precursor skills have predictive value beyond each other.

4. DISCUSSION

We examined the predictive value of the precursor skills of understanding indeterminacy and confounding for CVS skills 2 years later. Our analyses show that both indeterminacy and confounding are good and robust predictors of later CVS mastery overall, yet the predictive value of indeterminacy seems to be more pronounced. Particularly regarding the four individual sub-skills of CVS, indeterminacy stands out as a strong predictor even when controlling for systematic classroom differences and covariates (cognitive abilities and reading comprehension). In addition, we found differences in the two precursor skills' predictive strength for the four CVS sub-skills. Although indeterminacy seems to be especially predictive of children's later skills in identifying and interpreting controlled experiments, the understanding of confounding appears to be slightly more relevant for later skills in planning controlled experiments, and in detecting the inconclusiveness of confounded comparisons (representing the sub-skill *understanding*). These results might be slightly unexpected, since we assessed *confounding* as precursor by a task (see **Figure 1B**) based on a so-called choice-task (Bullock et al., 2009). Solving a choice-task requires an ability

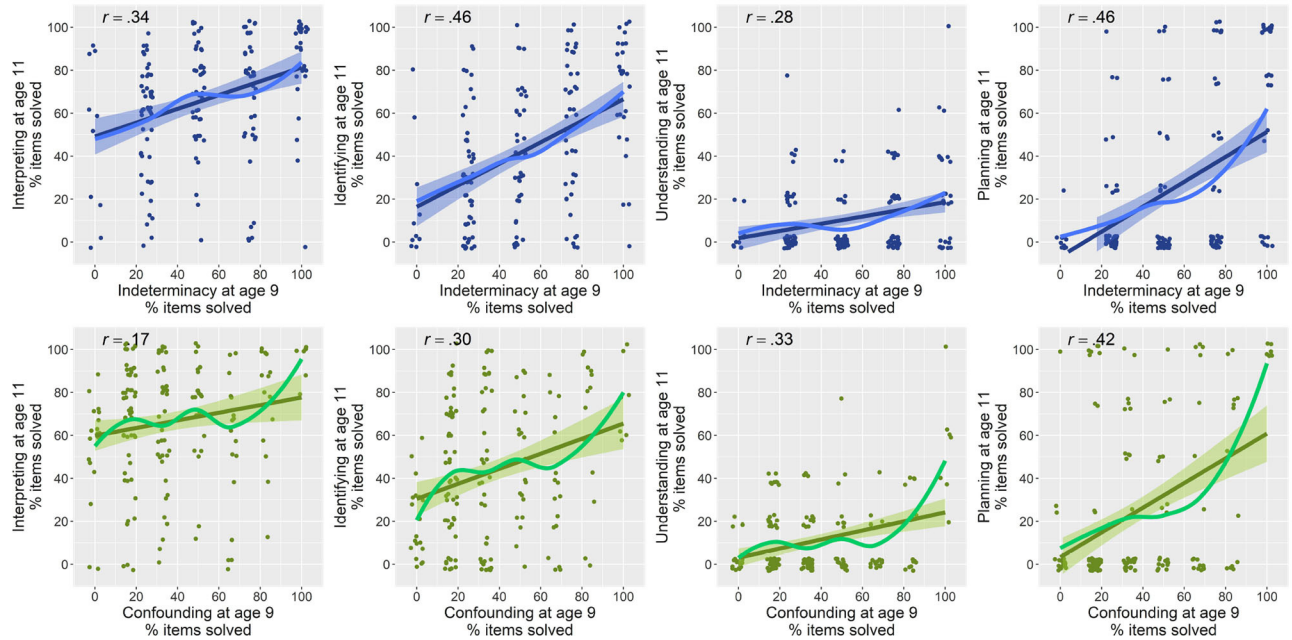


FIGURE 7 | Relations between children's precursory skills and later CVS skills. First row shows relations of indeterminacy with later skills, and second row for confounding. Pearson correlations between variables are indicated in upper left of plots.

TABLE 4 | Relative model fit and explained deviance for the 16 fitted general additive mixed models.

Model no.	Model description	AIC	BIC	% deviance explained
Model 1 int	Indeterminacy	1,601	1,611	10.8
Model 2 int	Confounding	1,615	1631	05.0
Model 3 int	Indeterminacy × confounding	1,604	1,623	11.6
Model 4 int	Indeterminacy × confounding, covariates, teacher	1,592	1,668	27.3
Model 1 ide	Indeterminacy	1,628	1,637	20.6
Model 2 ide	Confounding	1,650	1,667	10.9
Model 3 ide	Indeterminacy × confounding	1,626	1,644	23.1
Model 4 ide	Indeterminacy × confounding, covariates, teacher	1,589	1,677	46.4
Model 1 und	Indeterminacy	1,449	1,463	09.3
Model 2 und	Confounding	1,419	1,437	24.5
Model 3 und	Indeterminacy × confounding	1,420	1,445	24.9
Model 4 und	Indeterminacy × confounding, covariates, teacher	1,396	1481	42.5
Model 1 pla	Indeterminacy	1,678	1,693	22.3
Model 2 pla	Confounding	1,680	1,697	21.6
Model 3 pla	Indeterminacy × confounding	1,662	1,688	30.9
Model 4 pla	Indeterminacy × confounding, covariates, teacher	1,658	1,732	39.0

Covariates were reasoning ability and reading comprehension, including their interactions with indeterminacy and confounding on all four sub-skills of CVS as dependent variables (int, interpreting; ide, identifying; und, understanding; pla, planning). teacher indicates a random intercept across teachers.

which in our conceptualization receives the label *identifying* in fully developed CVS. A possible explanation for this finding might be that mastery of the tasks assessing later CVS skills requires skills going beyond *identifying*. The understanding of confounding shows stronger positive relations to all four sub-skills in the upper 20% of solved items (i.e., the associations are stronger for higher-scoring students). Hence, it shows stronger nonlinear relations than the understanding of indeterminacy. Finally, we found that the interactive predictive strength of the two precursor skills remains when taking into account children's more general cognitive development.

The pronounced yet not perfect predictive value of the two precursor skills indicates that although there is systematic stability in children's development of experimentation skills,

new individual differences arise during age 9–11. In addition, the predictive value of precursor skills differed between the four CVS sub-skills. One reason for the varying findings across CVS sub-skills could arise from different requirements that the two precursor skills and the four CVS sub-skills pose on children. Solving an item representing the understanding of indeterminacy requires connecting the outcomes with the initial variables and rethinking the experiment in a deductive manner. Solving items representing the understanding of confounding demands an inductive way of thinking. In the latter, children had to choose different variable levels from a given sample in order to solve the item and thus are able to gain new insights of the experiment. Additionally, the comprehension of confounding requires understanding a central mechanism of the CVS by keeping all additional variables beyond the focal variable constant.

There were several nonlinear relations between the variables in our study. An intuitive explanation for this finding is that variation in lower scores is mostly caused by guessing, whereas variation in higher scores is where mastery of the precursor skills and the later CVS sub-skills really comes into play. Consequently, the nonlinear nature of various relations in this study might be attributed to a methodological issue caused by multiple choice-tasks. However, this does not mean that modeling the nonlinearity in the relations is superfluous. To the contrary, when such nonlinear effects are neglected in statistical models, this will lead to underestimation of the real predictive value of one skill for another. Consequently, we suggest researchers who use multiple choice-tasks with limited range in their study to examine and consider modeling nonlinear relations as well.

Solving items that comprise the sub-skill *understanding* demands understanding the indeterminacy of confounded experiments. *Planning* requires the full understanding of CVS: Without applying the CVS correctly, no conclusive experiments can be planned consistently. These two skills seem to be closer in their mechanisms to the understanding of confounding. Peteranderl (2019) found that solving planning items showed a bimodal distribution. Either the children failed or they succeeded in most of the cases. This could explain the strong positive relation in the upper 20% of solving confounding items regarding the planning sub-skill. This explanation is related to findings by Schwichow et al. (2016a) and Schwichow et al. (2020). Those researchers found that the sub-skill *understanding* appears to be much more difficult than the other sub-skills even for secondary school students. They concluded that *understanding* might premise mastering the other sub-skills. A stronger predictive value of the precursor skill of confounding for detecting the inconclusiveness of confounded experiments and the positive relation for the upper 20% of all sub-skills support the conclusion of Schwichow et al. (2020) that full development in all other sub-skills yields further development in more difficult sub-skills, in particular *understanding*.

By contrast, the sub-skill *interpreting* demands the ability to interpret the result of a conclusive comparison correctly, whereas the sub-skill *identifying* demands the ability to select a conclusive comparison according to an initial hypothesis. Both sub-skills have been assessed within the same item contexts. First, children had to select a comparison that was conclusive, and second, they had to interpret this comparison. In our understanding of CVS, it is not feasible to separate these two sub-skills conceptually, as it was done in tasks such as the choice-task by Bullock et al. (2009), Osterhaus et al. (2020), and Piekny and Maehler (2013). According to Peteranderl (2019), it is important to have an overarching understanding of both sub-skills, and the ability to combine them into the same task is decisive for correct understanding of CVS. Nevertheless, there were many children who solved the first task (identifying) but failed in the second (interpreting) and vice versa. This finding indicates that the ability to apply one sub-skill of CVS does not necessarily cause correct application of the other sub-skill. This could explain the differing associations of the sub-skills interpreting and identifying with the two precursor skills and the stronger predictive value of the precursor of indeterminacy on both sub-skills.

The findings of our study add to the multidimensional perspective regarding the dimensionality of scientific reasoning (Mayer et al., 2014; Koerber et al., 2015; Edelsbrunner and Dablander, 2018). Our findings endorse the perspective of both CVS and its precursors, which just make up a part of scientific reasoning, as multidimensional constructs (Chen and Klahr, 1999; Schwichow et al., 2016a). If CVS or its precursor skills are assessed or modeled as unidimensional constructs, important differences in the predictive value of different skills and their predictive interplay would get lost (see also Edelsbrunner et al., in Press). In addition, some of the intercorrelations between the different CVS sub-skills were only moderate, even on the latent level, as indicated by a confirmatory factor analysis. Moderate intercorrelations even when controlling for measurement error variance via latent variable modeling indicate that in an overall score, a substantial amount of differential information would get lost. We therefore suggest that when overall scores of CVS or even broader scientific reasoning are used for analyses, researchers should check whether thereby relevant information that would be visible in more detailed sub-scores is lost. In addition, theoretical models of scientific reasoning should acknowledge that conceptualizing scientific reasoning as being strongly dominated by an encompassing core of understanding of the theory-evidence relation might not be a fruitful direction for research.

In line with the results of Schwichow et al. (2016a, 2020), we found that the sub-skill *understanding* shows the lowest performance of all four sub-skills, reflecting that these items are also the most challenging ones for younger children. We also had multiple children in our sample that could not solve *identifying*-items but were rather successful on *planning*-items, with the two latent variables behind these tasks correlating 0.56. Apparently, also these two sub-skills do not perfectly inform about each other, although they seem theoretically related. We believe that based on these results, researchers should critically consider under which circumstances, and for which research questions, assessing or modeling CVS, its precursor skills, or broader scientific reasoning as unitary constructs might be useful, or rather lead to a loss of useful information. We do not argue against the usage of overall scores or unidimensional models in general. If,

for example, researchers aim at modeling scientific reasoning as a competence that is equally represented by its constituents, an overall score or Rasch modeling might provide perfectly reasonable and informative approaches for achieving this aim. Our findings combined with prior studies (Edelsbrunner and Dablander, 2018; van Bork et al., 2019; VanderWeele and Batty, 2020) indicate that empirically, related questions might not be fully answered based on cross-sectional data, and that researchers should rather consider what models might be informative for the question at hand, pointing toward longitudinal examinations (or intervention studies) as a fruitful opportunity for examining dimensionality from a developmental perspective. A major reason is that longitudinal data provide the key for distinguishing between two reasons for an apparently (partially or more fully) unidimensional statistical structure: First, there can be a conceptual common core of scientific reasoning. Second, unidimensional structure can just appear because different skills influence each other over time; hence, unidimensionality arises that merely points toward developmental interplay (Van Der Maas et al., 2006).

Previous studies showed moderate to strong relations of facets of more general cognitive development with scientific reasoning (Morris et al., 2012; Piekny and Maehler, 2013; Mayer et al., 2014; van der Graaf et al., 2015, 2019; Osterhaus et al., 2017; Koerber and Osterhaus, 2019; Sande et al., 2019). Our results shows that the predictive value of both examined precursors on later mastery of CVS remains rather robust when taking into account general cognitive abilities and reading comprehension. Thus, new individual differences that arise when precursor skills develop into fully developed CVS cannot be explained by overarching more general cognitive development. Consequently, the development of CVS cannot be reduced to more general cognitive development, but also comprises domain-specific dynamics yielding CVS as an ability that is applicable in a domain-general manner similar to general reasoning, yet represents a distinct ability.

As a limitation of our study, it might be argued that posing the full CVS instrument from the second assessment on children twice, before and after the crucial developmental period, would have allowed more robust and additional insights. We agree with this point from a substantial perspective. However, posing a rather advanced instrument on children at age 9 might pose validity problems. The instrument might not work adequately in younger children, particularly with the *planning*-items that pose a rather complex task on children. We believe that it would be highly timely to develop an instrument that can be employed across a rather large period of childhood and beyond. In addition, we believe that validity issues, such as a lack of measurement invariance, might also arise because age 9–11 might not only be a period of increased quantitative development. Rather, this period might represent a time during which children more thoroughly restructure their knowledge base and skills regarding experimentation, representing qualitative development that would cause measurement issues when trying to assess the same skills over time.

Overall, our results shed more detailed light on previous findings (Sodian et al., 1991; Chen and Klahr, 1999; Bullock et al., 2009; Schwichow et al., 2016b) according to which children's understanding of CVS develops rapidly between age 9 and 11. Our longitudinal data suggest moderate individual stability in developing scientific reasoning skills. Earlier understanding of indeterminacy and confounding has clear yet limited predictive value for later mastery of CVS, going beyond that of more general cognitive development.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by ETH Zurich. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

SP was supported by an ETH Research Grant with the number ETH-23 15-1.

ACKNOWLEDGMENTS

We would like to thank Elsbeth Stern for contributing to the research question in this study and for providing feedback on an earlier draft, as well as Anne Deiglmayr, Lennart Schalk, and Ralph Schumacher, who have contributed to the relevant stages of the Swiss MINT Study.

REFERENCES

Bullock, M., Sodian, B., and Koerber, S. (2009). "Doing experiments and understanding science: development of scientific reasoning from childhood to adulthood," in *Human Development From Early Childhood to Early Adulthood: Findings From a 20 Year Longitudinal Study*, eds W. Schneider and M. Bullock (London: Psychology Press), 173–197.

Bullock, M., and Ziegler, A. (1999). "Scientific reasoning: developmental and individual differences," in *Individual Development From 3 to 12: Findings From the Munich Longitudinal Study*, eds F. E. Weinert, and W. Schneider (Cambridge: Cambridge University Press), 38–54.

Chen, Z., and Klahr, D. (1999). All other things being equal: acquisition and transfer of the control of variables strategy. *Child Dev.* 70, 1098–1120. doi: 10.1111/1467-8624.00081

Croker, S., and Buchanan, H. (2011). Scientific reasoning in a real-world context: the effect of prior belief and outcome on children's hypothesis-testing strategies. *Br. J. Dev. Psychol.* 29, 409–424. doi: 10.1348/026151010X496906

Dewey, J. (2002). *Logik: Die Theorie der Forschung*. Frankfurt am Main: Suhrkamp.

Dunn, T. J., Baguley, T., and Brunsden, V. (2014). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *Br. J. Psychol.* 105, 399–412. doi: 10.1111/bjop.12046

Edelsbrunner, P., and Dablander, F. (2018). The psychometric modeling of scientific reasoning: a review and recommendations for future avenues. *Educ. Psychol. Rev.* 31, 1–34. doi: 10.1007/s10648-018-9455-5

Edelsbrunner, P., Schalk, L., Schumacher, R., and Stern, E. (2018). Variable control and conceptual change: a large-scale quantitative study in elementary school. *Learn. Individ. Differ.* 66, 38–53. doi: 10.1016/j.lindif.2018.02.003

Edelsbrunner, P. A., Schalk, L., Schumacher, R., and Stern, E. (2015). "Pathways of conceptual change: Investigating the influence of experimentation skills on conceptual knowledge development in early science Education," in *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (Austin, TX), 620–625.

Edelsbrunner, P. A., Stern, E., and Schumacher, R. (in press). "Children's scientific reasoning skills in light of general cognitive development," in *The Cambridge Handbook of Cognitive Development*, eds O. Houdé and G. Borst (Cambridge: Cambridge University Press).

Gopnik, A., and Schulz, L. E. (2004). Causal learning across domains. *Dev. Psychol.* 40:162. doi: 10.1037/0012-1649.40.2.162

Groll, A., and Tutz, G. (2011). *Variable Selection for Generalized Additive Mixed Models by Likelihood-Based Boosting*. Munich: Ludwig-Maximilians-University.

Heller, K. A., and Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision: KFT 4-12+ R*. Beltz-Test.

Inhelder, B., and Piaget, J. (1958). *The Growth of Logical Thinking: From Childhood to Adolescence*. New York, NY: Basic Books. doi: 10.1037/10034-000

Klahr, D. (2000). *Exploring Science: The Cognition and Development of Discovery Processes*. Cambridge, MA: MIT Press.

Klahr, D., and Chen, Z. (2003). Overcoming the positive-capture strategy in young children: learning about indeterminacy. *Child Dev.* 74, 1275–1296. doi: 10.1111/1467-8624.00607

Klahr, D., Fay, A. L., and Dunbar, K. (1993). Heuristics for scientific experimentation: a developmental study. *Cogn. Psychol.* 25, 111–146. doi: 10.1006/cogp.1993.1003

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., and Sodian, B. (2015). The development of scientific thinking in elementary school: a comprehensive inventory. *Child Dev.* 86, 327–336. doi: 10.1111/cdev.12298

Koerber, S., and Osterhaus, C. (2019). Individual differences in early scientific thinking: assessment, cognitive influences, and their relevance for science learning. *J. Cogn. Dev.* 20, 510–533. doi: 10.1080/15248372.2019.1620232

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychol. Rev.* 96, 674–689. doi: 10.1037/0033-295X.96.4.674

Kuhn, D. (2007). Reasoning about multiple variables: control of variables is not the only challenge. *Sci. Educ.* 91, 710–726. doi: 10.1002/sce.20214

Kuhn, D. (2010). "What is scientific thinking and how does it develop?," in *Handbook of Childhood Cognitive Development, 2nd Edn*, ed U. Goswami (Chichester: Wiley-Blackwell). doi: 10.1002/9781444325485.ch19

Kuhn, D., Black, J., Keselman, A., and Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cogn. Instruct.* 18, 495–523. doi: 10.1207/S1532690XCI1804_3

Kuhn, D., Garcia-Mila, M., Zohar, A., Andersen, C., White, S. H., Klahr, D., et al. (1995). Strategies of knowledge acquisition. *Monogr. Soc. Res. Child Dev.* I–157. doi: 10.2307/1166059

Mayer, D., Sodian, B., Koerber, S., and Schwippert, K. (2014). Scientific reasoning in elementary school children: assessment and relations with cognitive abilities. *Learn. Instruct.* 29, 43–55. doi: 10.1016/j.learninstruc.2013.07.005

Morris, B. J., Croker, S., Masnick, A. M., and Zimmerman, C. (2012). "The emergence of scientific reasoning," in *Current Topics in Children's Learning and Cognition*, ed H. Kloos (London: IntechOpen), 61–82.

Osterhaus, C., Koerber, S., and Sodian, B. (2017). Scientific thinking in elementary school: children's social cognition and their epistemological understanding promote experimentation skills. *Dev. Psychol.* 53:450. doi: 10.1037/dev0000260

Osterhaus, C., Koerber, S., and Sodian, B. (2020). The science-p reasoning inventory (SPR-I): measuring emerging scientific-reasoning skills in primary school. *Int. J. Sci. Educ.* 1–21. doi: 10.1080/09500693.2020.1748251

Peteranderl, S. (2019). *Experimentation skills of primary school children* (Ph.D. thesis). ETH Zurich, Zurich, Switzerland.

Piekny, J., and Maehler, C. (2013). Scientific reasoning in early and middle childhood: the development of domain-general evidence evaluation, experimentation, and hypothesis generation skills. *Br. J. Dev. Psychol.* 31, 153–179. doi: 10.1111/j.2044-835X.2012.02082.x

Piéraut-Le Bonniec, G. (1980). *The Development of Modal Reasoning: Genesis of Necessity and Possibility Notions*.

Sande, E., Kleemans, M., Verhoeven, L., and Segers, P. (2019). The linguistic nature of children's scientific reasoning. *Learn. Instruct.* 62, 20–26. doi: 10.1016/j.learninstruc.2019.02.002

Schalk, L., Edelsbrunner, P. A., Deiglmayr, A., Schumacher, R., and Stern, E. (2019). Improved application of the control-of-variables strategy as a collateral benefit of inquiry-based physics education in elementary school. *Learn. Instruct.* 59, 34–45. doi: 10.1016/j.learninstruc.2018.09.006

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Dev. Psychol.* 32:102. doi: 10.1037/0012-1649.32.1.102

Schauble, L., Klopfer, L. E., and Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *J. Res. Sci. Teach.* 28, 859–882. doi: 10.1002/tea.3660280910

Schneider, W., Schlagmüller, M., and Ennemoser, M. (2017). *LGVT 5-12+: Lesegeschwindigkeits-und-Verständnistest für die Klassen 5-12: Manual*. Hogrefe.

Schwichow, M., Christoph, S., Boone, W., and Härtig, H. (2016a). The impact of sub-skills and item content on students' skills with regard to the control-of-variables strategy. *Int. J. Sci. Educ.* 38, 216–237. doi: 10.1080/09500693.2015.1137651

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., and Härtig, H. (2016b). Teaching the control-of-variables strategy: a meta-analysis. *Dev. Rev.* 39, 37–63. doi: 10.1016/j.dr.2015.12.001

Schwichow, M., Osterhaus, C., and Edelsbrunner, P. A. (2020). The relation between the control-of-variables strategy and content knowledge in physics in secondary school. *Preprint retrieved from edarxiv.org/82xdk*. doi: 10.1016/j.cedpsych.2020.101923

Siegler, R. S., Liebert, D. E., and Liebert, R. M. (1973). Inhelder and piaget's pendulum problem: teaching preadolescents to act as scientists. *Dev. Psychol.* 9:97. doi: 10.1037/h0035073

Siler, S., Klahr, D., Magaro, C., Willows, K., and Mowery, D. (2010). "Predictors of transfer of experimental design skills in elementary and middle school children," in *International Conference on Intelligent Tutoring Systems* (Berlin: Springer), 198–208. doi: 10.1007/978-3-642-13437-1_20

Siler, S. A., and Klahr, D. (2012). "Detecting, classifying and remediating children's explicit and implicit misconceptions about experimental design," in *Psychology of Science: Implicit and Explicit Reasoning*, eds R. W. Proctor and E. J. Capaldi (New York, NY: Oxford University Press), 137–180.

Simonsohn, U. (2019). *Interaction Effects Need Interaction Controls*. Blog post on Datacolada.

Sodian, B., Zaitchik, D., and Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Dev.* 62, 753–766. doi: 10.2307/1131175

Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Dev.* 51, 1–10. doi: 10.2307/1129583

van Bork, R., Rhemtulla, M., Waldorp, L. J., Kruis, J., Rezvanifar, S., and Borsboom, D. (2019). Latent variable models and networks: Statistical equivalence and testability. *Multivariate Behav Res*. doi: 10.1080/00273171.2019.1672515. [Epub ahead of print].

van der Graaf, J., Segers, E., and Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: dynamic assessment of the control of variables strategy. *Instruct. Sci.* 43, 381–400. doi: 10.1007/s11251-015-9344-y

van der Graaf, J., Segers, E., and Verhoeven, L. (2016). Discovering the laws of physics with a serious game in kindergarten. *Comput. Educ.* 101, 168–178. doi: 10.1016/j.compedu.2016.06.006

van der Graaf, J., van de Sande, E., Gijsel, M., and Segers, E. (2019). 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.* 41, 1119–1138. doi: 10.1080/09500693.2019.1594442

Van Der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., and Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychol. Rev.* 113:842. doi: 10.1037/0033-295X.113.4.842

VanderWeele, T. J., and Batty, C. J. (2020). On the dimensional indeterminacy of one-wave factor analysis under causal effects. *arXiv [Preprint]*. arXiv:2001.10352.

Wagensveld, B., Segers, E., Kleemans, T., and Verhoeven, L. (2015). Child predictors of learning to control variables via instruction or self-discovery. *Instruct. Sci.* 43, 365–379. doi: 10.1007/s11251-014-9334-5

Wilhelm, P., and Beishuizen, J. (2003). Content effects in self-directed inductive learning. *Learn. Instruct.* 13, 381–402. doi: 10.1016/S0959-4752(02)00013-0

Wood, S. N. (2017). *Generalized Additive Models: An Introduction With R*. Chapman and Hall; CRC Press. doi: 10.1201/9781315370279

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Dev. Rev.* 27, 172–223. doi: 10.1016/j.dr.2006.12.001

Zimmerman, C., and Croker, S. (2013). Learning science through inquiry. na.

Zimmerman, C., and Klahr, D. (2018). "Development of scientific thinking," in *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, Vol. 4, ed J. Wixted (Hoboken, NJ: Wiley), 1–25. doi: 10.1002/9781119170174.epcn407

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Peteranderl and Edelsbrunner. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership