

Relational Reasoning and Visuospatial Tools:
Unlocking STEM Learning and Reasoning

By

Elena R. Leib

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Silvia A. Bunge, Chair
Professor Jan M. Engelmann
Professor Michelle H. Wilkerson
Professor Priti Shah

Spring 2024

Relational Reasoning and Visuospatial Tools:
Unlocking STEM Learning and Reasoning

Copyright 2024

by

Elena R. Leib

Abstract

Relational Reasoning and Visuospatial Tools:
Unlocking STEM Learning and Reasoning

by

Elena R. Leib

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor Silvia A. Bunge, Chair

Humans excel at detecting patterns in information, abstracting rules, and making inferences. Underlying these skills is relational reasoning: the cognitive ability to identify and map abstract, generalizable relations between pieces of information. Though this powerful ability supports higher-order cognition, it can also be a processing bottleneck when the complexity of information is too high for our limited cognitive resources. My dissertation explores this dynamic—relational reasoning as both a cognitive tool and bottleneck for learning and reasoning—and how we overcome cognitive limitations by offloading relations to external spatial representations, such as visuospatial tools (e.g., graphs and diagrams). I focus much of my work on STEM outcomes in children because many important concepts and skills in these disciplines are relational in nature and difficult for students to learn, making them an ideal testbed for these empirical questions.

In Chapter 1, I take a high-level view of the relationship between reasoning and education and review evidence that education hones reasoning ability. I find significant evidence that the protracted, immersive experience of formal schooling taxes, and therefore improves, general reasoning skills, such as relational reasoning.

In Chapter 2, I establish that there is a unique role for relational reasoning in learning that is distinct from other cognitive skills. In two empirical studies, I use the case of fraction learning to investigate the main executive functions involved in fraction processing, and then show that relational reasoning predicts fraction performance over and above these other strong domain-general predictors.

In Chapter 3, I investigate how we learn to offload relations to physical space. In an empirical study with the Tsimane', an indigenous farmer-forager people from the Amazon basin, I find that individuals spontaneously offload to-be-remembered relations to space, including individuals who report no formal schooling and are not literate. This study demonstrates that offloading

relations to external representations is part of a foundational cognitive toolkit and is separate from the regular use of visuospatial tools.

Finally, in Chapter 4, I show that scaffolding relational reasoning during learning can improve understanding, focusing on the case of graph comprehension. I first propose a relational reasoning perspective on the difficulties that children and adults have with graph comprehension. Then, I empirically test aspects of this approach in a preliminary intervention study that manipulates the extent to which relational reasoning is engaged during a lesson on interpreting graphs of linear functions. I find that having students focus on patterns—both visual and conceptual—and make comparisons supports learning.

Taken together, this body of work simultaneously contributes to our understanding of the role of relational reasoning in higher-order cognition and to applications of relational reasoning for improving STEM education. In particular, my research provides a generative framework for identifying the STEM content that students may find especially difficult, as well as for informing the design of pedagogical approaches for helping students overcome these obstacles.

Dedication

*To all of my family and friends
who love me, support me, and
help me grow every day.*

Acknowledgments

This journey has been as much about the people as it has been about the research, if not more, and I have many people to thank.

Silvia, thank you for taking a chance on me and for your mentorship and support over these 6 years! I feel incredibly grateful to have had the opportunity to grow as a researcher with you and be part of your lab. I am also deeply grateful that working with you brought me to Berkeley, both the city and the university. Thank you for the opportunity to develop and follow my ideas, and for connecting me with colleagues who complement your mentorship. Thank you for sticking by me, believing in me, helping me learn how to protect my time, and always helping me move forward in the program. It has been a joy and honor to work with you.

Steve, thank you for also taking a chance on me and sending me an email in early May 2022 asking if I would be interested in going to Bolivia. That email truly changed the course of my grad school career in a most exciting and meaningful way, more than I could have ever imagined! Having the opportunity to work with you, your lab, and visit San Borja was a deeply formative experience for me both as a researcher and more broadly as an individual. Thank you!

Thank you to my amazing committee for their feedback, insights, and support over the years. Jan, your love of research is contagious! I enjoy discussing research ideas with you, and our semesterly coffee chats always left me feeling inspired, motivated, and more confident in myself. Thank you for your enthusiasm for my ideas and general excitement for me at every stage of grad school. Michelle, thank you for expanding my thinking in so many ways and for guiding me into new literatures. You opened my eyes to thinking about epistemic tools, which fundamentally shifted the way I approach my research. Further, thank you for welcoming me into your writing group during lockdown and creating such a supportive community of researchers. Priti, thank you for your excitement about my research and encouraging me at every step of the way. This was especially meaningful when I was feeling uncertain and down on my research ideas. After meetings with you I always felt a boost of motivation, confidence, and enthusiasm, in addition to having many new ideas for my projects.

In addition to my advisors and committee, I was lucky to work with two more wonderful mentors. Ariel, I really can't imagine where I'd be in grad school without you! I am grateful that I started working on that first project with you my first year and that I've gotten to continue working with you ever since. I have learned so much from you—from how you wrote data cleaning and analysis scripts, to your clear thinking, excellent writing, and efficient use of time, just to name a few. Thank you for all of your support, many hours of Zoom calls, and care. You have been there for me through all the ups and downs. Miriam, I am grateful that I had the opportunity to work with you and learn from you! I especially appreciate the time you spent with me editing and re-writing

parts of our manuscript over Zoom. I learned a great deal from watching your process and hearing you think aloud. Thank you for your time and mentorship.

When I started grad school, I could never have imagined all the amazing, deep friendships that I would build!

Monica, I knew from the day I met you when you hosted me at interviews that I wanted to be your labmate and your friend. And lucky for me, I have gotten to be both! You are always there for me, no questions ask, and you get me like few others do. I'm endlessly grateful to have you as my well buddy and my heart person, and to journey together on the same emotional wavelength.

Willa, developing QuACK with you and teaching it together every year has been one of the biggest sources of growth, pride, and joy for me throughout grad school. I love brainstorming with you, thinking big, and fleshing out new ideas—there's nothing like that shared excitement when we know we are on to something. Thank you for helping me to grow as a teacher and for believing in me. And for being the best climbing buddy! I'm glad we have a new project together.

Roya, your love, joy, care, and curiosity are infectious, and you have an incredible way of making even difficult times bright. Thank you for being a spark of joy and fun from the day I met you at interviews, for your incredible memory of details about people as well as about every paper you've ever read, and for always having the perfect snacks and sharing them with me.

Thank you to the Bunge Lab and the Colala Lab for lab lunches, interesting discussions, and good laughs. Thank you Holly for always being there for a chat or a hug. I love our lunches and stopping by your desk, and I'm grateful we've had each other from the beginning. Thank you to the Developmental Area for being a warm, supportive, and fun home base and to the Psychology Department as a whole, I could not have asked for a better community of students, staff, and faculty. I am also grateful for our Social Dev Prosem class from Fall 2019 and to Ari and Jan who were teaching it for their first time. That was a special learning community, thank you for expanding my thinking around the role of society and cultural context in development. Thank you to my 2018 cohort, I'm lucky to have gone through grad school with such a brilliant, motivated, accomplished, and kind group of people.

A meaningful part of my time as a grad student has been as a teacher and mentor, and I am grateful to all of my students for coming to class with open minds, getting excited about the material, believing in themselves, and supporting me as I grew as a teacher. Special thanks to my former QuACK students for your interest and dedication and for co-creating an inclusive and welcoming space with Willa and me. Thank you for giving us the opportunity and freedom to explore what we love and do what we care about. We could not have done it without you!

Thank you to all my teachers and mentors over the years, starting from childhood. I especially want to thank my math and science teachers, who created welcoming and engaging

environments for me to learn, and my psychology professors and research mentors and advisors. Thank you to all the children and adults who participated in my studies and to my research assistants over the years, including Hana Massab, Roshni Sarathy, Royalle Hurney, and Emily Kleinfelder. Thank you to my funding sources, including the Berkeley Fellowship, Jacobs Foundation, and NSF. And a very special thanks to all of the dedicated folks who have developed R, RStudio, and the Tidyverse packages, and to the whole R community for creating all the tools, packages, and help pages that I have used to wrangle, analyze, and visualize my data. I cannot thank you enough—without you and these tools this research and these beautiful plots would not be possible! And more than just code, you have created an amazing and supportive community of learners. Thank you!

And of course I wouldn't be anywhere without my incredible family.

Mom and Dad, thank you for supporting me and encouraging me in everything I do. Mom, thank you for sharing with me your interest in and compassion for people, your love of teaching and connecting, and your joy and appreciation for the beauty of the world around us. I'm so glad I bring you joy and nachas. Dad, thank you for sharing with me your computational and analytical thinking skills, your attention to detail, your love of patterns and numbers, and baking. I am grateful you suggested I take a computer science class in college! You both are wonderful role models and have taught me how to build community wherever I go.

Josh, I feel so lucky that in you I have a wonderful brother and a best friend. I admire your curiosity, insightfulness, and intentionality, and am grateful that I can go to you for anything, from advice to a hearty laugh. I love spending time with you!

Grandma and Grandpa, thank you for everything you have done for our family and for teaching me the meaning of hard work, dedication, resilience, and perseverance. Grandma, I cherish our time together. Thank you for always asking what is new and then following up about it. I feel so lucky that I can confide in you. I also appreciate when you remind me that sometimes you just need to tell yourself that you are fine and get things done. Grandpa, I think about you and miss you all the time. Thank you for always encouraging my interests in math, science, and computers. I know you would be so proud of me!

Mom Mom, thank you for your warmth, love, and care. I treasure our time together, whether it is in Philadelphia, Florida, Pittsburgh, or on the phone. I can always count on you to tell me a good story or new fact I didn't know, make me laugh, and help me finish a crossword, to name just a few. I love that we make each other feel seen and special. Pop Pop, I wasn't old enough to know you well, but luckily I've gotten to hear many stories. I know I share with you a love of connecting with people, teaching, and the beach. I often think about you when I'm in the classroom and how proud you'd be to see me teaching, too.

Omead, thank you for being you. I feel incredibly lucky and grateful that I met you when I did. You add so much to my life every day and are one of the most supportive and caring people I know. Thank you for being a calming energy and my adventure buddy, for helping me take fun breaks so I don't get burned out, making me delicious food, giving me huge hugs, and making me laugh. I am so lucky to have you by my side.

To all my family and friends, I'm certain I could write a whole other dissertation just on how incredible you all are! Thank you for your endless love, care, support, and laughs. For introducing me to new activities and ideas, pushing me to grow, and providing the warm environment to sustain that growth. For being with me through thick and thin, laughs, cries, rants, and everything in between. For keeping me grounded and for helping remind me of who I am and who I want to become. I wish I could fully communicate even a fraction of how much I love you and how special you are to me. Thank you!

A little reflection for my past self, my future self, and anyone reading this:

The path is always winding. You don't always know where you're going or how you'll get there, but you are stronger and more resilient than you can ever know. Keep following what excites you and you'll always be doing something that you love.

Table of Contents

Abstract	1
Acknowledgments	ii
Table of Contents	vi
Introduction	1
Relational reasoning and relational complexity	1
A mechanism for learning and a cognitive bottleneck	3
Reasoning and relational offloading with visuospatial tools	4
Relational reasoning and STEM education	6
The present work	7
Chapter 1. Education hones reasoning ability	8
1.1 Abstract	8
1.2 Introduction	8
1.3 Effects of formal education on reasoning	9
1.4 Reasoning programs designed by researchers	9
1.5 Courses that tax reasoning skills	10
Preparing for a law school entrance exam	10
Emerging findings	12
1.6 Broader considerations	13
1.7 Conclusion	13
Chapter 2. Relational reasoning is distinct from other domain-general cognitive processes	15
2.1 General Introduction	15
2.2 Testing the whole number interference hypothesis: contributions of inhibitory control and whole number knowledge to fraction understanding	15
Abstract	15
Introduction	16
Methods	21
Results	28
Discussion	36
2.3 Relational thinking: An overlooked component of executive functioning	44
Abstract	44
	vi

Introduction	44
Method	48
Results	53
Discussion	61
Chapter 3. Spontaneous and strategic relational offloading to physical space	64
3.1 General Introduction	64
3.2 Indigenous Amazonians spontaneously use space to offload cognitive demands	65
Abstract	65
Introduction	65
Experimental paradigm	66
Participants spatially organized cards to strategically represent relevant information	68
Shape of card layouts varied within and between conditions	70
Use of spatial strategies increased after the first trial	71
Case studies show change in strategy	72
Discussion	73
Materials and Methods	74
Chapter 4. Scaffolding relational reasoning: A promising approach for promoting graph comprehension	79
4.1 Abstract	79
4.2 Introduction	79
Relational reasoning and relational complexity	81
4.3 A relational reasoning perspective on graph comprehension	81
Relational reasoning in the cognitive processing of graphs	82
Relational reasoning and the levels of graph comprehension difficulty	83
Relational complexity in graphical displays	85
Implications for graph pedagogy	86
4.4 Preliminary study	87
Methods	89
Results	99
4.5 Discussion	102
Limitations and future directions	105
Conclusion	107
General Discussion	108
Summary and theoretical implications	108
Pedagogical implications	109

Conclusion	110
References	111
Appendices	134
Appendix A: Supplemental Materials for Chapter 2	134
Appendix B: Supplemental Materials for Chapter 3	135
Appendix C: Supplemental Materials for Chapter 4	157

Portions of this dissertation appear in the following articles:

Chapter 1

Bunge, S. A., & Leib, E. R. (2020). How Does Education Hone Reasoning Ability? *Current Directions in Psychological Science*, 29(2), 167–173.

<https://doi.org/10.1177/0963721419898818>

Chapter 2

Section 2.2

Leib, E. R., Starr, A., Younger, J. W., Project iLead Consortium, Bunge, S. A., Uncapher, M. R., & Rosenberg-Lee, M. (2023). Testing the whole number interference hypothesis: Contributions of inhibitory control and whole number knowledge to fraction understanding. *Developmental Psychology*, 59(8), 1407–1425.

<https://doi.org/10.1037/dev0001557>

Section 2.3

Starr, A., Leib, E. R., Younger, J. W., Project iLead Consortium, Uncapher, M. R., & Bunge, S. A. (2023). Relational thinking: An overlooked component of executive functioning.

Developmental Science, 26(3), e13320. <https://doi.org/10.1111/desc.13320>

Chapter 3

Leib, E. R., Bunge, S. A., & Piantadosi, S. T. (*under review*). Indigenous Amazonians spontaneously use space to offload cognitive demands.

Chapter 4

Leib, E. R., Wilkerson, M. H., Shah, P., & Bunge, S. A. (*in prep*). Scaffolding relational reasoning: A promising approach for promoting graph comprehension

Introduction

In our day-to-day lives, we often need to reason about complex information across various domains. For example, we choose what foods we should eat, make plans for how to spend our days, and think about pressing issues in society, such as whether we can change the trajectory of climate change. One feature that these diverse examples have in common is that they require the reasoner to think about and coordinate many different *relations* between multiple pieces of information. In the case of planning a day, a reasoner must consider various dimensions of each activity on her to-do list, such as how long each will take to complete and how urgent and important it is, and integrate those dimensions with additional factors, such as how much energy she has, enjoyment, the weather outside, and so on. Underlying these thinking skills is relational reasoning: the domain-general cognitive ability to map abstract relations between pieces of information. This capacity is considered core to human cognition and has been shown to support abstraction, inference, generalization, analogical reasoning, and fluid reasoning (Gentner, 2003; Halford et al., 2010; Hofstadter, 2001). However, it can also be a processing bottleneck when the complexity of information is too high for our finite cognitive resources.

My dissertation explores this dynamic—relational reasoning as both a cognitive tool and bottleneck for learning and reasoning—and how we overcome cognitive capacity limits by offloading relations to external spatial representations, such as graphs and diagrams. A second aim is to further bridge the relational reasoning literature with education and pedagogy, exploring ways that a relational reasoning lens can improve instruction in complex domains. In the following four chapters, I investigate 1) whether education hones reasoning, 2) whether there is a unique role for relational reasoning in learning, over and above other cognitive skills, 3) the cognitive origins of relational offloading, and 4) whether scaffolding relational reasoning during learning can improve understanding. In much of this work, I use STEM content and outcomes as a testbed for these questions. Before getting into the details of these chapters, I first start by introducing the concepts of relational reasoning and relational complexity. Next, I overview how relational reasoning is a powerful mechanism for learning and how it is constrained by prior knowledge and finite cognitive resources. I then apply the lens of relational reasoning to visuospatial tools (e.g., graphs and diagrams) to explain how they could be used to offload some of the relational demand during reasoning. After, I contextualize relational reasoning and visuospatial tools in STEM fields and make a case for why these areas are a generative testbed for theories about relational reasoning. Finally, I overview each chapter of the dissertation and the research questions they address.

Relational reasoning and relational complexity

Relational reasoning is the ability to identify and map abstract, generalizable relations between two or more objects, pieces of information, or representations. A relation is a predicate that takes two arguments (Gentner, 1983). For example, *larger*(elephant, dog), which can be read as “The elephant is larger than the dog” and *cause*(kick, injury) as “The kick caused the injury.” Thus, relational reasoning can be thought of as a function that takes the arguments as input, operates on them (i.e., maps relations between them), and outputs a new piece of information, such as a

generalization, which is more abstract than the inputs. This process has been investigated empirically and modeled extensively (e.g., Bunge et al., 2005; Falkenhainer et al., 1989; Halford et al., 2012; Hummel & Holyoak, 2005; Johnson-Laird, 2010). Relational representations are constructed in working memory, where arguments (e.g., elephant and dog) dynamically bind into roles in the relation, which means that the same arguments can later bind to other relations (e.g., *older*(elephant, dog)) or fill different slots in the same relation (e.g., *larger*(dog, mouse)), depending on the context (Halford et al., 2010; Hummel & Holyoak, 2005). Converging evidence from cognitive, education, and neuroscientific studies show that relational reasoning develops slowly over childhood and into early adolescence, reaching adult-like levels around 11 to 12 years old (Andrews & Halford, 2002; Crone et al., 2009; Jablansky et al., 2016; Wendelken et al., 2017). Various features affect the relational reasoning process, two of which are the focus here: the nature and number of arguments entering into the relation.

The nature of the arguments—whether they are objects or other relations—affects the *order* of the relation. A first-order relation takes objects as arguments, such as the *larger*(elephant, dog) example above (Gentner, 1983). It has been demonstrated that even infants are sensitive to first-order relations of cause and effect, simple spatial relations (e.g., above, below), and quantitative relations (e.g., more than, less than) (Goswami, 2001; Penn et al., 2008). Another common first-order relation is object similarity, a *same* relation that takes two objects and compares their identity by mapping the relationship between their features. For example, infants are able to abstract an ABA pattern from hearing sets of phonemes, such as ‘la-di-la’ and ‘bo-ta-bo’ (R. L. Gómez & Gerken, 2000; Marcus et al., 1999). These abstractions are perceptually bound because the arguments are percepts (e.g., *same*(sound_pos1, sound_pos3)). Alexander (2016) suggests that perceptually bound first-order relational reasoning is what helps humans recognize patterns in the continuous flow of information that floods our senses, allowing for percepts to be coupled with concepts to help make sense of the world around us.

However, relational reasoning is not limited to operating just on objects: we can also reason about relations between relations. Termed higher-order relational reasoning, these relations take at least one other relation as an argument (Gentner, 1983). Gentner (1983) gives the example of *cause*[*strike*(x,y), *collide*(y,z)], which can be read as “X strikes y, causing y to collide with z.” Another example of higher-order relations is relational similarity, which includes analogy. Like object similarity it uses the *same* relation, but instead of objects or percepts as arguments now the arguments are relations, allowing for comparisons between the relational structures of two different representations (Gentner, 2003). For example, in saying that the hydrogen atom is like the solar system a reasoner maps similarity between the relational structure of the solar system (e.g., *revolve_around*(planets, sun)) to that of a hydrogen atom (e.g., *revolve_around*(electrons, nucleus)). In this way, higher-order relational reasoning is role-based: it is not the attributes nor surface-level features of the objects themselves that are being compared, but rather the roles of the objects and relationships among them (Gentner, 1983; Halford et al., 2010; Hummel & Holyoak, 2005).

In addition to the nature of arguments that enter into a relation, the number of arguments, or *arity*, is also important. For example, *larger*(elephant, dog) is a binary relation because it has slots

for two arguments. Addition, on the other hand, represents a ternary relation that takes three arguments (e.g., $2 + 3 = 5$ can be written as *addition*(2, 3, 5)). Relational complexity is the processing and working memory load imposed by the number of variables that are being related. The greater the number of variables, the greater the relational complexity and cognitive load. Thus, relational reasoning is a cognitively demanding process that quickly becomes more taxing as the number of items to be related increases.

A mechanism for learning and a cognitive bottleneck

Relational reasoning serves as a general learning mechanism that drives the acquisition of increasingly complex, abstract, and relational knowledge throughout childhood and adulthood. Gentner (2003) and Goswami (2001) both posit that the relational processes present in human infants are the same as those that support higher-order relational reasoning in children and adults and that these processes underlie humans' incredible learning abilities (see also Cattell, 1943; Hebb, 1942). Relational reasoning scaffolds learning by taking input from prior knowledge or current representations in the environment and outputting information that is more abstract—e.g., a pattern, category, generalization, or induction—and may be novel to the reasoner. In turn, this newly learned knowledge can either become the input for another round of relational processing, continuing to build more abstract knowledge, or stored in memory for later use. Thus, relational reasoning serves as an iterative process that supports learning.

However, these powerful relational processes are constrained by limitations in prior knowledge and cognitive resources, which I argue can result in bottlenecks in learning and reasoning. Regarding prior knowledge, if an individual does not yet know a relation, they cannot use that concept in their reasoning. For example, if a math student does not yet know the *part-whole* relation, they will have difficulty reasoning about fractions. Further, prior knowledge has been shown to affect how individuals represent problems and process relations (e.g., Chi et al., 1981; Gentner, 2003; Gick & Holyoak, 1980). For example, younger children tend to make more feature-based comparisons whereas older children make more relational comparisons (Gentner, 1988; Rattermann & Gentner, 1998), and novices are more likely to attend to surface-level features in problems whereas experts extract the relational structure (Chi et al., 1981).

In addition to prior knowledge, cognitive load can also constrain relational reasoning, as it scales with the number of variables to be related. For example, consider a situation where you are trying to decide what food to make for a party. If cost is your only consideration, the decision may be relatively straightforward. However, as you consider more variables—the time it will take to cook a dish, people's dietary restrictions, what vegetables are in season, what you enjoy cooking, etc.—you can feel the difficulty of the problem increasing. In fact, it has been estimated that four variables is the maximum number that can be integrated in a single processing step without invoking other strategies, and that as processing limits are reached, speed and accuracy of reasoning declines (Andrews & Halford, 2002; Halford et al., 1998).

Luckily, when relational overload occurs, it typically does not terminate processing, but rather prompts the reasoner to adopt a strategy for reducing relational demand. Halford et al. (1998) discuss two such strategies. The first is to reduce a relation into a simpler concept. For example,

the density of a substance is the amount of mass per each unit of volume. Instead of reasoning about density as a proportion, which is often difficult, this relation can be simplified by using the single number for density, making it one variable, or unary as Halford et al. (1998) call it. In fact, this strategy is quite common; we regularly simplify all kinds of proportions to one variable, such as speed (miles per hour) and standard deviation (distance from the mean per data point). Though this reduction can be beneficial for processing, it also has its downsides because the reasoner temporarily loses access to the component relations and variables that make up the concept. A second strategy is to segment the complex relation into less complex steps and process them serially rather than in parallel. However, a third strategy, which is not discussed frequently enough in the relational reasoning literature, is what I term relational offloading: the option to offload relations to external visuospatial representations.

Reasoning and relational offloading with visuospatial tools

Humans have invented a variety of spatial tools for reducing the cognitive demands of day-to-day tasks, including writing systems, calendars, graphs, and diagrams. Here, I focus on the latter two—graphs and diagrams—which were invented to aid in reasoning by externally representing the relations in information in physical space (Hegarty, 2011; Tversky, 2011). Using these tools can assist the reasoner either as part of the process of uncovering patterns in information or to help communicate them to others. Individuals use these tools to offload demands onto physical space, extending cognition and freeing up cognitive resources to think more abstractly, reason and remember more effectively, and even make discoveries (Bauer & Johnson-Laird, 1993; Clark & Chalmers, 1998; Hegarty, 2010; Kirsh, 2010; Risko & Gilbert, 2016; Tversky, 2015). Learning how to use and create them is critical for children and adults alike. Graphs and diagrams are used across many different disciplines in the humanities, social sciences, and hard sciences. Further, literacy with these tools is increasingly more important to be an engaged citizen as well as to understand, protect, and have agency over one's own data.

Despite a non-trivial literature on the cognitive science of these visuospatial tools and implications for design and instruction (e.g., Bauer & Johnson-Laird, 1993; Franconeri et al., 2021; Hegarty, 2011; Shah et al., 2005; Shah & Hoeffner, 2002; Tversky, 2011), few researchers have investigated the role of relational reasoning in processing these external representations, or how their use affects the complexity of problem solving. I propose that relational reasoning is involved in three ways. First, these tools are designed to communicate relations, so relational reasoning is likely needed to uncover the pattern of relations between the variables and interpret the message being communicated. Second, the way that these tools represent the relations likely also demands relational reasoning. Diagrams and graphs are abstract simplifications of the world. They depict information about variables and the relations between them via spatial and visual features, such as points, lines, arrows, proximity, and color (Hegarty, 2011; Tversky, 2011). Thus, to make sense of these representations, the reasoner must map the marks on the page to the real-world information and relations that they represented. I propose that this mapping process involves relational reasoning. Third, the information gleaned from these external representations is typically integrated into larger reasoning problems. Graphs and diagrams are rarely used in isolation; more often, they are used to provide evidence in support of an argument or conclusion.

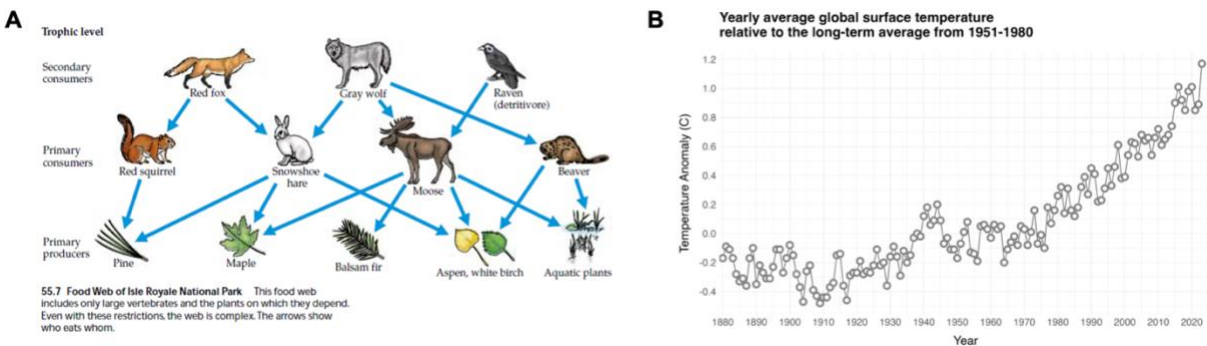


Figure 0.1. Examples of a (A) diagram and (B) graph. (A) Image retrieved from <https://prior.allenai.org/projects/diagram-understanding>. (B) Data retrieved from NASA’s Goddard Institute for Space Studies, <https://climate.nasa.gov/vital-signs/global-temperature/>.

Task analyses with two examples, one of a diagram and one of a graph, will further elucidate these three roles for relational reasoning.

Consider the diagram of an ecosystem’s food web (Figure 0.1A). Arrows and position on the page (higher or lower) are used to communicate who eats who. The arrows represent the relation *eats* (e.g., *eats*(fox, squirrel)) and the three position levels on the page (top, middle, or bottom) are mapped to three different trophic levels. To make sense of this diagram, a reasoner must make these mappings between the spatial relations and conceptual relations. Now imagine you are using this diagram to help you reason about what will happen to the gray wolf and red fox populations if a disease has killed most of the balsam fir trees in the area. Instead of having to keep in mind and integrate the relations between 12 animal and plant species, these relations are offloaded to the diagram, and you can use your eyes to look from each species of interest down the food chain. Using transitive inference, a form of relational reasoning (Halford et al., 2010; Wendelken & Bunge, 2010), you may conclude that since the gray wolf population would be more negatively affected by the Balsam fir disease than the red fox because gray wolves eat moose, which eats balsam fir, but none of the fox’s prey eat balsam fir.

Next, consider the graph showing the relation between year and global temperature anomaly relative to a long-term average of temperatures from 1951-1980 (Figure 0.1B). Each point represents a year, and the height of the point represents the temperature anomaly that year. Though there is no trend line on the graph—the black line simply connects the points in sequential order—you likely see an “upward” trend, which you map to meaning that temperature anomaly is increasing over the years. Now imagine you are using this graph to predict how warm 2025 will be. Integrating across all these points and considering the upward trend in addition to apparent variability year-to-year, you estimate that the temperature anomaly will be approximately +1.1 degrees Celsius.

Ironically, perhaps, these very tools that are intended to help facilitate identifying and communicating relations are often difficult for novices to learn (e.g., Diezmann, 2000; Friel et al.,

2001; Glazer, 2011; Shah & Carpenter, 1995; Shah & Hoeffner, 2002). In particular, their deep relational structure—both in what they aim to communicate and how the information is represented—is difficult to extract if you have not explicitly learned the mapping conventions or practiced extracting information from these representations. As illustrated in the examples in Figure 0.1, a reasoner needs to learn that “up” in the food web is mapped to being “higher” in the food chain, but that “up” in the graph is mapped to a greater value of the y variable. Relatedly, little is known about how we come to use space in these systematic ways to offload relations, a question I investigate in Chapter 3. Perhaps we learn these strategies in school; alternatively, systematically using space to offload relations may be a fundamental part of our cognitive toolkit, which is then enhanced by schooling and learning these formal tools. In addition to investigating this basic research question, I also focus on the use of these formal visuospatial tools in STEM domains in Chapter 4.

Relational reasoning and STEM education

In many of the studies in this dissertation, I focused on STEM learning and reasoning. Researchers have stressed an important role for relational reasoning in scientific reasoning and inquiry (Dumas, 2017; Klahr et al., 2013; Resnick et al., 2017), education and academic achievement (Dumas et al., 2013; Richland & Simms, 2015; Vendetti et al., 2015), and STEM education specifically (Alexander, 2017; DeWolf et al., 2015; Kalra et al., 2020; Miller Singley & Bunge, 2018; Murphy et al., 2017). In Chapter 2, I extend this research to show that the role of relational reasoning is separable from other domain-general cognitive processes. Researchers have also stressed an important role for spatial thinking and reasoning in these fields (e.g., Atit, Power, et al., 2020; Atit, Uttal, et al., 2020; Ishikawa & Newcombe, 2021; Taylor et al., 2023). In Chapter 4, I investigate the relations between relational reasoning, spatial thinking, and STEM education, focusing on relational offloading and scaffolding relational reasoning during learning.

STEM fields provide a rich testbed for the dynamics of relational reasoning and relational offloading for two main reasons. First, the content of STEM fields is highly relational (Chi et al., 1981; DeWolf et al., 2015; Goldwater & Schalk, 2016; Kalra et al., 2020; Richland et al., 2012). From learning about food webs to density, fractions, algebra, and chemical interactions, these fields focus on understanding the relations between variables. Second, higher-order relational reasoning underlies many of the skills needed to carry out the core epistemic practices of these disciplines. For example, analogy has been cited as one of the most important and widely used cognitive tools for scientific innovation (Klahr et al., 2013). Analogies can be used for generating new theories, hypotheses, and scientific explanations. More broadly, scientific reasoning, which makes up an important part of everyday reasoning (Klahr et al., 2013; Kuhn, 2010), involves uncovering hidden causal relationships between variables and an outcome, coordinating theory and evidence, finding abstract patterns, making comparisons, and building mental models—all of which require higher-order relational reasoning to map relations and extract the relational structure (Johnson-Laird, 2010; Kuhn et al., 1992). Visuospatial tools are also central to epistemic practices of these fields, including food web diagrams, molecular structures, free-body diagrams, the periodic table, and graphs of functions in math, and graphs of data in science and statistics (Collins & Ferguson, 1993; Hegarty, 2011; Taylor et al., 2023; Tversky, 2015). As described above,

many of these tools are inherently relational, both in the type of information that is represented and how it is represented.

An additional reason for my research interest in STEM fields is that they are notoriously difficult for students to learn, and there are many knowledge-based, sociocultural, and systemic barriers to entry. Relatedly, many students often simply do not enjoy these fields, likely in part due to how they are taught. I propose that a relational reasoning approach to learning in STEM domains has the potential to improve pedagogy. Therefore, focusing on STEM content to investigate relational reasoning questions is mutually beneficial and contributes to the relational reasoning literature, expands what we know about learning in these STEM fields, and ultimately could make these domains more engaging and broadly accessible. The ways in which relational reasoning capacity can be scaffolded or leveraged to support STEM education is a nascent, yet active, area of research, and there are still many open questions.

The present work

The overarching goals of my dissertation work are to investigate the dynamics of relational reasoning as both a cognitive tool and bottleneck, and to explore how offloading relations to external representations can help overcome cognitive limitations. In many studies, I use STEM domains as a testbed for these questions, which in turn can inform STEM pedagogy from a relational reasoning perspective.

In Chapter 1, I take a high-level view of the relationship between reasoning and education and review evidence that, in addition to the fact that reasoning supports academic achievement, the experience of education also hones reasoning. In Chapter 2, I establish that relational reasoning is separable from other cognitive processes that are involved in learning math. In two empirical studies, I investigate the main executive functions that are involved in fraction learning and understanding, and then show that relational reasoning predicts performance on fraction problems over and above these other strong domain-general predictors. In Chapter 3, I investigate how we learn to offload relations. For this empirical study, I worked with the Tsimane', an indigenous farmer-forager people from the Amazon basin of Bolivia who live in a non-industrialized society and often have minimal levels of formal education and literacy. I tested two leading hypotheses about relational offloading to physical space: either it is a specific strategy that is culturally transmitted, or it is broadly available as part of our cognitive toolkit, allowing individuals to innovate ad hoc relational offloading strategies even in novel contexts. Finally, in Chapter 4, I investigate whether scaffolding relational reasoning during learning can help students overcome cognitive bottlenecks. To do so, I first offer a relational reasoning perspective on graph processing, comprehension difficulties, and pedagogy. I then conduct an intervention study that begins to test this approach by manipulating the extent to which relational reasoning is engaged during a lesson on graph comprehension. Taken together, this program of research provides a generative framework for identifying the STEM content that students may find especially difficult, as well as for informing the design of pedagogical approaches for helping students overcome these obstacles.

Chapter 1. Education hones reasoning ability

1.1 Abstract

A belief about education that dates back several millennia is that, in addition to imparting specific facts, it hones general cognitive abilities that can be leveraged for future learning. However, this idea has been a source of heated debate over the past century. Here, we focus on the question of whether and when schooling hones reasoning skills. We point to research demonstrating cognitive benefits of both broad and specific educational experiences. We then highlight studies that have begun to elucidate underlying mechanisms of learning. Given our society's substantial investment in education, it behooves us to understand how best to prepare individuals to participate in the modern workforce and tackle the challenges of daily living.

This chapter contains previously published material from the following work:

Bunge, S. A., & Leib, E. R. (2020). How Does Education Hone Reasoning Ability? *Current Directions in Psychological Science*, 29(2), 167–173. <https://doi.org/10.1177/0963721419898818>

1.2 Introduction

A common assumption is that education prepares students for the challenges that lie ahead: that beyond imparting specific facts, it hones general cognitive skills like reasoning, which can be deployed in new contexts to solve novel problems. This assumption dates back at least as far as Plato, who theorized around 380 B.C.E. that training on math could transfer to reasoning about politics and ethics (Burnyeat, 2000). He espoused what came to be known in the late 18th century as the doctrine of formal discipline, which holds that studying rigorous subjects disciplines the general faculties of the mind. This doctrine is at the core of many of our educational institutions to this day. But is this assumption correct? There has been a heated debate for over a century, on and off, regarding the extent to which learning transfers across contexts and tasks (e.g., Judd, 1908; Redick, 2019; Singley & Anderson, 1989; Thorndike & Woodworth, 1901).

Here, we make the case that the immersive, multifaceted, protracted experience of formal schooling taxes, and therefore hones, general cognitive skills that can support learning across multiple domains (Ceci, 1991). We argue that it is necessary to probe the cognitive and neural mechanisms of learning more deeply (Gabrieli, 2016; Lindenberger et al., 2017) in order to address the question of *what* transfers, *how*, and for *whom* (Barnett & Ceci, 2002; Judd, 1908; Katz et al., 2018). We propose that probing how the subtle learning effects that are evident on the order of weeks to months can provide mechanistic insights regarding the types of learning that take place across multiple years of schooling.

We focus below on the question of whether and how schooling can sharpen the capacity for reasoning. Common measures of reasoning, including those used in IQ tests, require *relational* reasoning, or the ability to compare or integrate the relations among disparate pieces of information. Relational reasoning is conceptualized as an all-purpose cognitive ability that enables us to compare the magnitudes of two fractions, derive logical conclusions from a set of premises, understand the analogies used to teach scientific concepts, and more (Alexander, 2016; Holyoak, 2012). Indeed, there is a large body of evidence that relational reasoning is an important

predictor of scholastic achievement and other important life outcomes (Dumas & Dong, 2022; Goldwater & Schalk, 2016). Although the abstract tasks used to probe relational reasoning predict learning across multiple domains, there is a fair degree of pessimism regarding whether students hone this purportedly domain-general ability through instruction, or whether they instead learn only to reason about specific content matter (see Nisbett et al., 1987).

Below we briefly review the evidence that schooling can indeed hone relational reasoning (for in-depth reviews, see Ceci, 1991; Ritchie & Tucker-Drob, 2018). We then turn our attention to recent investigations probing *how* it does so. The three classes of studies discussed below have investigated the effects of 1) formal education writ large, or of pursuing a specific academic discipline, 2) curricula designed by researchers to explicitly teach reasoning skills, and 3) existing courses that were not developed by researchers, but whose effects on reasoning have been studied.

1.3 Effects of formal education on reasoning

How can we study whether schooling hones reasoning, short of randomly assigning children either to attend school or not to? One of several clever ways to get at this question leverages the fact that children of the same age can be in different grades, as a result of which it is possible to tease apart schooling-related and age-related improvements in cognitive performance (Morrison et al., 2019). A large study adopting this approach in over 12,000 Israeli children revealed a large effect of schooling on tests of reasoning across three grade levels (Cahan & Cohen, 1989). A more recent study showed better reasoning among first-graders than kindergarteners of roughly the same age (Q. Zhang et al., 2019).

Integrating the results of numerous and diverse studies, Ritchie and Tucker-Drob (2018) concluded that IQ scores (which are heavily based on reasoning test performance) rise 1-5 points for every additional year of education. Other studies have distinguished between the types of reasoning emphasized in different disciplines. For example, one study showed that students in the social sciences improved more at statistical and methodological reasoning over the course of their undergraduate training than did those in the natural sciences or humanities (Lehman & Nisbett, 1990). This line of work suggests that students specializing in different fields learn to reason in different ways (Nisbett et al., 1987).

1.4 Reasoning programs designed by researchers

In a second class of studies, researchers have developed and evaluated courses targeting various types of reasoning skills through explicit instruction. One example is a 10-lesson curriculum built on the observation that diverse reasoning tasks have a common element of comparing objects or relations among objects (Klauer et al., 2002)—what we refer to as relational reasoning. A review based on 74 studies involving nearly 3600 children and adolescents suggests that this curriculum works as intended: after roughly 5 weeks of reasoning instruction, children showed improvements on other tests of reasoning that lasted for months (Klauer & Phye, 2008). Another example is a *gist reasoning* program designed to teach strategies for “glean[ing] deep meaning from texts through analysis and synthesis of information, inference of abstract concepts, prediction of outcomes, and relating what is presented in text to one’s own background

knowledge” (Gamino et al., 2014). This program consists of 8-12 sessions administered over 1-2 months. It has been shown to improve gist reasoning across multiple populations (Chapman & Mudar, 2014), including adolescents from a wide range of socioeconomic backgrounds (Gamino et al., 2014). The results of these curricula are promising, although it remains to be seen whether the effects on reasoning are evident across a broad range of tasks.

1.5 Courses that tax reasoning skills

In a third class of studies, researchers have examined the effects of completing a specific course that exists ‘in the wild’ rather than having been developed in the lab. This scientific approach falls somewhere between the broad, real-world ‘intervention’ of schooling as a whole and focused, well-controlled interventions: it has real-world relevance and is experimentally tractable. In the 1980s, this approach was adopted to test for effects of computer programming instruction on reasoning; findings were mostly negative or inconclusive (Salomon & Perkins, 1987). There are many plausible explanations, including the possibility that such far transfer is not possible (Singley & Anderson, 1989). However, newer studies have the advantage of several additional decades of refinement of cognitive theory and methodology and are well-positioned to revisit this question. For example, we can ask: does completing a specific course promote the application of newly learned rules and strategies (Halpern, 2001), change the way we represent a problem (Cetron et al., 2019), and/or improve the efficiency of domain-general cognitive processes that undergird reasoning (Guerra-Carrillo & Bunge, 2018)?

Preparing for a law school entrance exam

In our lab, we have leveraged several methods to evaluate whether and how preparation for the Law School Admission Test (LSAT) hones relational reasoning. The LSAT was our curriculum of choice because a full two-thirds of the test focuses explicitly on teaching strategies for different types of reasoning; the remaining third focuses on reading comprehension. In our first study, we compared pre-law students who had just enrolled in a 3-month LSAT test preparation course with a passive control group of well-matched pre-law students. The course included 70 hours of explicit reasoning instruction and practice.

Using functional magnetic resonance imaging (fMRI), we found that taking the LSAT course was associated with changes in brain regions associated with reasoning (Figure 1.5.1A; Mackey et al., 2012, 2013). Specifically, we found changes in measures of brain anatomy and brain function that are thought to reflect the strength of communication between regions in a network. Thus, the course had an impact on the neural machinery that has been implicated in a wide range of reasoning tasks. These findings lend credence to the idea that learning to solve LSAT problems could lead to improvements on other reasoning tasks.

We also found changes in behavioral performance and brain activation measured with fMRI while participants performed a relational reasoning task (Figure 1.5.1B). These abstract problems bear no resemblance the text-based LSAT problems (see Figure 1.5.2A and C), but they both require reasoning about relations between different pieces of information. Compared with the controls, the LSAT students showed a larger improvement in both accuracy and response times. Moreover, fMRI analyses indicated that they relied less on the *dorsolateral prefrontal cortex*, a brain region

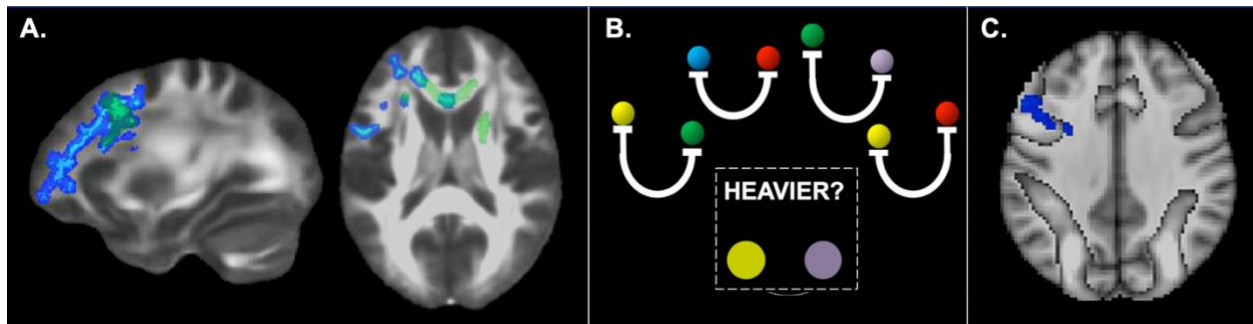


Figure 1.5.1. A) Two slices of the brain, one showing the left hemisphere from the side (with the front of the brain on the left of the image) and the other showing the brain from above (front of the brain at the top of the image). The green and blue clusters show white matter changes associated with LSAT course completion, as measured with diffusion tensor imaging. Effects were observed in regions that have been implicated in reasoning, including left prefrontal cortex and the bundle of fibers that connects the left and right prefrontal cortex (the anterior corpus callosum), as well as parietal cortex (not shown). These results support the hypothesis that reasoning instruction led to increased white matter connectivity. (Adapted from Mackey et al., 2012). B) Sample problem from a transitive inference task that measures relational reasoning ability. In this problem, participants have to encode that the purple ball is heavier than the green one and that the green ball is heavier than the yellow one in order to determine that the purple ball is heavier than the yellow ball. After LSAT training, participants completed this task more accurately and more quickly—even though this task looks nothing like the LSAT problems. C) Slice of the brain from above, showing changes in brain activation associated with reasoning instruction, as measured with fMRI. The gray areas outlined in black are the regions engaged during performance of the transitive inference task shown in panel B. Shown in blue is a region in dorsolateral prefrontal cortex that exhibited a decrease in activation after taking the LSAT course. A decrease in activation suggested that participants in the LSAT group were able to perform the transitive inference task more efficiently. (Adapted from Mackey et al., 2015).

that is consistently engaged during performance of challenging tasks, suggesting that they found the task easier (Figure 1.5.1C; Mackey et al., 2015). This study provides evidence of moderate transfer of learning from a real-world reasoning curriculum to a laboratory-based test of reasoning.

By contrast, we did not find transfer to any of four standardized cognitive measures, including two other tests of relational reasoning. This discrepancy may highlight the need for a better taxonomy of reasoning tasks, detailing the cognitive processes that underlie each of them, and how much overlap there is in terms of the cognitive demands of the curriculum and the outcome measures (Klauer & Phye, 2008). We have adopted eyetracking with a view to pinpointing cognitive processes that are impacted by an intervention, and ultimately understanding why transfer to other tasks succeeds or fails. Eye movements, which are among the fastest movements that the human body makes, reflect shifts in attention that are associated with thought processes. The location and sequence of eye gaze fixations provide us with a rich source of data that goes beyond what we can get with accuracy or response times and that complements brain imaging data. In conjunction with behavioral data, eyetracking has the potential to provide novel insights regarding learning (Eckstein et al., 2017).

In a follow-up to the LSAT study, we used eyetracking to probe transfer effects more deeply. We randomly assigned pre-law students to take a 6-week online course focused on either the Analytical Reasoning section of the LSAT (text-based reasoning problems; Figure 1.5.2A) or the

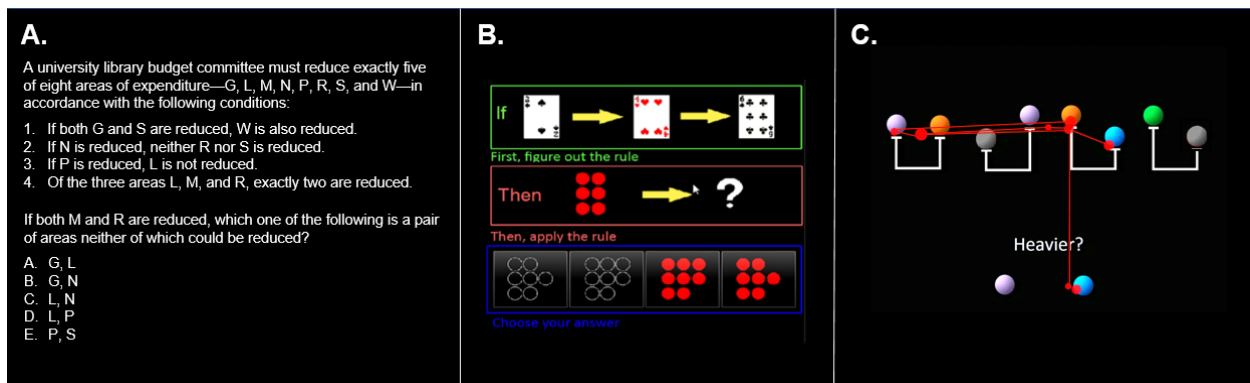


Figure 1.5.2. A) Sample LSAT Analytical Reasoning question from the Law School Admission Council website (www.lsac.org). The correct answer is C. B) One of the four reasoning tests that were completed by participants before and after the online LSAT course (either Analytical Reasoning or Reading Comprehension) to test for transfer from the online course to reasoning skills. In this task, the participant must first induce the rule from the cards, and then apply it to a novel problem. (Adapted from Guerra-Carrillo & Bunge, 2018). C) A participant’s eye gaze pattern during completion of a transitive inference problem just before the response was made. On average, participants made 23 eye movements in the 7 seconds it took to solve these problems. Participants in the Analytical Reasoning group improved in speed on this task. Examining the reasoning group’s gaze patterns revealed that this improvement was not due to becoming faster at initially identifying the relevant relations, but rather, participants spent less time looking at the relevant relations after identifying them, suggesting they improved in the efficiency of their relational thinking. Further, the degree to which an individual participant improved on the task was most strongly related to the magnitude of change on this ocular measure (Guerra-Carrillo & Bunge, 2018).

Reading Comprehension section of the LSAT (Guerra-Carrillo & Bunge, 2018). We observed an effect of reasoning instruction on a composite score of four relational reasoning tests that were visually distinct from the LSAT problems (e.g., Figure 1.5.2B), but not on other cognitive measures. These results support *moderate* transfer of learning.

Reasoning instruction also led participants to perform the transitive inference problems more quickly, while maintaining high accuracy. We analyzed changes in how participants’ eyes moved around the computer screen as they examined these problems. The eye gaze data allowed us to adjudicate between several possible mechanisms of learning by providing evidence that increased efficiency of relational thinking was the most important change (for details, see Figure 1.5.2C). The eye gaze data revealed that the *individuals who showed the biggest improvements* on the transfer task showed the biggest change in our eyetracking measure of relational thinking. Thus, we gained insights that could be used to evaluate whether a particular program taxes the target cognitive processes, whether it might be necessary to extend or modify the program to maximize behavioral benefits, and what works best for whom.

Emerging findings

Other exciting findings are beginning to emerge. For example, Adam Green and colleagues have as-yet unpublished data showing that a year-long high school geoscience course that taxes spatial reasoning is associated with improvements on both spatial and non-spatial reasoning tests, alongside changes in brain activation in regions implicated in relational reasoning. Another study shows that the brains of undergraduates who have developed expertise in mechanical engineering register correspondences between objects that, despite being visually dissimilar,

share deeper physical properties (Cetron et al., 2019). These new approaches help to demystify when and how transfer of learning occurs. In the United States, current educational standards emphasize scientific reasoning; this presents an opportunity to investigate whether or to what extent standards-aligned curricula hone reasoning across scientific disciplines, and beyond.

1.6 Broader considerations

One critical question is whether there is a particular window in development when schooling is most likely to hone general cognitive skills. Improvements in reasoning and in the underlying anatomical connections are most pronounced during the elementary school years (Wendelken et al., 2017); thus, this may be a time when this neurocognitive system is particularly malleable. There is a great need for further research examining schooling effects on child brain development (see Brod et al., 2018).

Another critical question is how long one could reasonably expect schooling effects to last. We anticipate long-lasting benefits only if students continue to leverage the skills they have honed, or if there has been a fundamental change in the way they represent information (Lövdén et al., 2010). Earlier studies suggested that children’s IQ scores drop over long summer holidays, and that this ‘summer slide’ is particularly large for students from socioeconomically disadvantaged backgrounds (Ceci, 1991). These results warrant replication, as they have important implications for reducing the achievement gap.

We have posited that the time in life when many individuals are at their peak level of cognitive functioning is while they are still in school, practicing thinking skills and acquiring knowledge at break-neck speed. In a large online study, we found differences not only in the *level* of cognitive performance across educational brackets, but also—more compellingly—in the age at which *peak* cognitive performance was observed *within* each educational bracket (Guerra-Carrillo et al., 2017). While we did not have the opportunity to follow the participants over time, it is intriguing that peak functioning within each group was observed around the typical time of completion of that degree. We posit that schooling effects could explain why cognitive performance tends to rise quickly during childhood and adolescence, peak in the early twenties, and then decline slowly throughout adulthood (McArdle et al., 2002).

1.7 Conclusion

In closing, there is evidence that the process of educating ourselves can equip us to reason about novel problems. This piece is by no means a comprehensive review; there are certainly many counter-examples of approaches that have been ineffective (e.g., see Singley & Anderson, 1989). The most promising curricula are likely those that aim to *teach for transfer* by encouraging deep understanding, explicitly teaching thinking skills using a variety of examples, and drawing attention to the structural features of a problem (Chi & VanLehn, 2012; Halpern, 2001; Willingham, 2008). Here, we call for deeper exploration of such curricula, both via replication studies and individual differences analyses. Ultimately, we need to understand why transfer of learning may be broader under some circumstances, and for some students, than for others.

Given our significant investment in education as individuals and families, and as a society, we must continue to assess the claim that Plato made 2400 years ago, asking ourselves: how can we best prepare students for future learning (Bransford & Schwartz, 1999)? Good reasoning skills are needed to master new job requirements as needed to keep pace with rapid technological advances (World Economic Forum, 2018), and to make sound decisions in all aspects of our lives. Finding ways to more effectively cultivate reasoning could therefore have profound and far-reaching consequences for society at large.

Chapter 2. Relational reasoning is distinct from other domain-general cognitive processes

2.1 General Introduction

Despite a consensus in the relational reasoning literature that this cognitive process is important for learning and academic achievement (e.g., Alexander, 2017; Dumas et al., 2013; Richland & Simms, 2015; Vendetti et al., 2015), the role of relational reasoning has not often been studied in the context of other mid-level, domain-general cognitive processes that support goal-directed behavior. These executive functions (EFs) are also important for learning, reasoning, and academic achievement (Best et al., 2011; Diamond, 2013; Lawson & Farah, 2015; Richland & Burchinal, 2013; Rose et al., 2011), and some have argued that relational reasoning and these EFs are one in the same (e.g., Martínez et al., 2011). In this chapter, I use the case of fraction understanding to test whether there is a unique role for relational reasoning. In section 2.2, I establish that two EFs, working memory and inhibitory control, are important for fraction understanding over and above whole number knowledge. In section 2.3, I examine the pattern of relations among these EFs and relational reasoning and then test whether relational reasoning explains additional variance in fraction understanding over and above the EFs. Indeed, I find that relational reasoning explains unique variance in fraction understanding, suggesting that it is separable and distinct from other EFs.

This chapter contains previously published material from the following work:

Leib, E. R., Starr, A., Younger, J. W., Project iLead Consortium, Bunge, S. A., Uncapher, M. R., & Rosenberg-Lee, M. (2023). Testing the whole number interference hypothesis: Contributions of inhibitory control and whole number knowledge to fraction understanding. *Developmental Psychology*, 59(8), 1407–1425.

<https://doi.org/10.1037/dev0001557>

Starr, A., Leib, E. R., Younger, J. W., Project iLead Consortium, Uncapher, M. R., & Bunge, S. A. (2023). Relational thinking: An overlooked component of executive functioning. *Developmental Science*, 26(3), e13320.

<https://doi.org/10.1111/desc.13320>

2.2 Testing the whole number interference hypothesis: contributions of inhibitory control and whole number knowledge to fraction understanding

Abstract

The present study tests two predictions stemming from the hypothesis that a source of difficulty with rational numbers is interference from whole number magnitude knowledge. First, inhibitory control should be an independent predictor of fraction understanding, even after controlling for working memory. Second, if the source of interference is whole number knowledge, then it should hinder fraction understanding. These predictions were tested in a racially and socioeconomically diverse sample of US children (N=765; 337 female) in grades 3 (ages 8-9), 5 (ages 10-11), and 7 (ages 12-13) who completed a battery of computerized tests. The fraction comparison task included problems with both shared components (e.g., $3/5 > 2/5$) and distinct components (e.g., $2/3 > 5/9$), and problems that were congruent (e.g., $5/6 > 3/4$) and incongruent (e.g., $3/4 > 5/7$) with whole number knowledge. Inhibitory control predicted fraction comparison performance over and above working memory across component and congruency types. Whole

number knowledge did not hinder performance and instead positively predicted performance for fractions with shared components. These results highlight a role for inhibitory control in rational number understanding and suggest that its contribution may be distinct from inhibiting whole number magnitude knowledge.

Introduction

Rational numbers, especially fractions, are a persistent stumbling block in the elementary and middle school mathematics curriculum. Identifying the root causes of fraction difficulties is a vital step in developing instructional programs to strengthen rational number knowledge. One proposed source of challenge in understanding fractions is that many properties of whole numbers do not hold for rational numbers, and the tendency to inappropriately apply properties of whole numbers when working with rational numbers has been termed whole number bias (Ni & Zhou, 2005). Among whole numbers, for example, larger numerals always denote larger quantities (e.g., $9 > 3$), but this is not true for fractions (e.g., $1/9 < 1/3$), and lower performance on fraction comparisons where whole number knowledge contradicts the appropriate rational number response is well-documented (Braithwaite & Siegler, 2018; DeWolf & Vosniadou, 2015; Fazio et al., 2016; D. M. Gómez & Dartnell, 2018; Meert et al., 2010; Miller Singley & Bunge, 2018; Obersteiner et al., 2013). Implicit in this literature is what we term the *whole number interference hypothesis*, that whole number knowledge interferes with rational number processing, leading to the observed performance decrements. In the present study, we use an individual differences design to test two predictions that stem from this hypothesis: 1) inhibitory control supports fraction performance, and 2) whole number magnitude knowledge hinders performance.

With respect to the first prediction, if interference resolution is crucial for fraction understanding, we would expect inhibitory control—the capacity to withhold prepotent responses and resolve interference—to be a strong and independent predictor of fraction performance. Prior work has shown that executive functions, including inhibitory control and working memory—the capacity to maintain and manipulate information—support academic outcomes (Best et al., 2011; Lawson & Farah, 2017; Rose et al., 2011). In fact, working memory is the most robust predictor of mathematical outcomes in children (Bull & Lee, 2014; Peng et al., 2016). Although prior rational number studies have examined the contributions of these executive functions separately (Bailey et al., 2014; D. M. Gómez et al., 2015; Jordan et al., 2013; Matthews et al., 2016; Siegler et al., 2012), there is little evidence of whether inhibitory control contributes to fraction outcomes when controlling for working memory. Establishing that inhibitory control uniquely contributes to fraction outcomes over and above working memory would bolster the claim that rational number difficulties stem, at least partly, from *interference* effects.

With respect to the second prediction, if the source of interference is more specifically *whole number magnitude knowledge*, we would expect individuals with better understanding of whole numbers to, perhaps paradoxically, perform worse on fraction tasks. This prediction therefore is a strong test of the whole number interference hypothesis, as it runs counter to the large body of research demonstrating whole number magnitude knowledge supports general math achievement (Schneider et al., 2017; Smedt et al., 2013), assessments of which often include rational number items. While whole number knowledge can refer to magnitudes, arithmetic

operations, or place values, in the current study, we use the term *whole number knowledge* to mean knowledge about the magnitudes of whole numbers.

Eliciting whole number interference in fraction comparison

The standard approach to eliciting whole number interference in fraction comparison is to manipulate fraction pairs' congruency with whole number knowledge. This manipulation takes a different form for fraction pairs that share components (i.e., either same denominator or same numerator) versus pairs that have distinct components (i.e., neither the numerator nor denominator is the same). Among the simpler problems with *shared components*, same denominator pairs are congruent with whole number knowledge because the fraction with the larger numerator has the larger magnitude (e.g., $4/5 > 3/5$). Thus, it is possible to arrive at the correct response by using whole number comparison to select the fraction with the larger numerator. In contrast, same numerator pairs are incongruent with whole number knowledge because the fraction with the smaller denominator has the larger magnitude (e.g., $3/5 > 3/8$). In these cases, using whole number comparison to select the fraction with the larger denominator would lead to the incorrect response. Consistent with the whole number interference hypothesis, accuracy is typically higher for congruent compared to incongruent shared component problems (Braithwaite & Siegler, 2018; DeWolf & Vosniadou, 2015; Fazio et al., 2016; D. M. Gómez & Dartnell, 2018; Meert et al., 2010; Miller Singley & Bunge, 2018).

Among the more challenging problems with *distinct components*, congruent pairs are typically defined as pairs in which the fraction with the largest components also has the larger magnitude (e.g., $5/6 > 3/4$); whereas in incongruent pairs, the fraction with the largest components has the smaller magnitude (e.g., $2/3 > 5/9$; but see Ischebeck, Schocke, and Delazer (2009) for an exception). The assumption underlying this manipulation for distinct components is that congruent pairs should be easier than incongruent pairs because comparing either the numerators or the denominators will lead to the same correct (or incorrect) response, an effect generally born out in the literature (D. M. Gómez et al., 2015; D. M. Gómez & Dartnell, 2018). In this study, we seek to use individual differences in inhibitory control and whole number knowledge to examine how whole number interference contributes to fraction comparison performance.

Executive functions and fraction understanding

Domain-general cognitive capacities, such as executive functions (EFs), have a well-established contribution to academic outcomes. Among the canonical EF constructs of working memory, inhibitory control, and cognitive flexibility (Diamond, 2013), working memory is one of the most robust predictors of mathematical outcomes (Friso-van Den Bos et al., 2013; Peng et al., 2016). The contributions of inhibitory control and cognitive flexibility are less clear, as many studies do not find relations with these constructs and general measures of math achievement (K. Lee & Bull, 2016; K. Lee & Lee, 2019; Van Dooren & Inglis, 2015). One possible explanation for the lack of consistent effects for inhibitory control is that it may be involved primarily in mathematical domains that require resolving interference from prior knowledge.

A similar phenomenon has been found in science education, where inhibitory control predicts learning when students must undergo conceptual change but not learning of factual information (Bascandziev et al., 2018). In fact, rational numbers have been cited as a paradigmatic example of conceptual change because learners must expand their understanding of number from discrete and countable whole numbers to continuous and dense rational numbers (Carey, 2011; Vamvakoussi & Vosniadou, 2004). These dueling conceptions of number make rational numbers an ideal testbed for clarifying the role of inhibitory control in math outcomes.

A growing list of studies have examined correlations between inhibitory control and rational number comparison, for both fractions and decimals (Abreu-Mendoza et al., 2020; Avgerinou & Tolmie, 2019; D. M. Gómez et al., 2015; Ren & Gunderson, 2021). Yet, these studies did not explicitly measure working memory. Further, although a handful of studies have looked at the role of working memory in fraction learning, it is more frequently considered a control variable rather than a variable of interest (Bailey et al., 2014; Jordan et al., 2013; Siegler et al., 2012; Starr et al., 2023). One recent study involving adolescents did consider how inhibition and visual working memory relate to three factors derived from performance on a standardized math task: basic arithmetic, rational number transformations, and fraction arithmetic and basic algebra (Abreu-Mendoza et al., 2018). Both inhibition and working memory correlated with the factors, but no regression analysis directly contrasted the contributions of these skills (Abreu-Mendoza et al., 2018). Thus, while both EF components contributed to rational number outcomes, this study did not establish whether the contribution of inhibitory control is distinct from that of working memory.

One study to date has considered contributions of all three canonical EF dimensions in a rational number comparison task. Coulanges et al. (2021) found that in college students, both working memory and inhibitory control independently predicted decimal comparison performance in counterintuitive pairs in which the decimal with the larger numerical value contained fewer digits (e.g., $0.8 > 0.27$). Interestingly, inhibitory control only predicted performance on these whole number knowledge conflicting problems, whereas working memory was also related to performance on problems that did not involve conflict (e.g., $0.80 > 0.27$), and cognitive flexibility did not contribute to either form of decimal comparison. These results highlight the centrality of working memory and inhibitory control for rational number understanding, but also point to distinctions between them. Working memory may support performance across a range of problem types, whereas inhibitory control may be especially important when intuitions based on prior knowledge, such as whole number knowledge, interfere with the correct response. In addition, the contributions of these cognitive factors may vary in younger participants who have less experience with rational numbers.

In the current study, we considered whether inhibitory control has a specific influence on incongruent fraction comparisons, which contradict prior whole number knowledge, or contributes comparably to both congruent and incongruent comparisons. If inhibitory control is needed exclusively for overcoming whole number magnitude knowledge, then we would expect to find an influence of inhibitory control only on incongruent pairs. On the other hand, if inhibitory control is also needed to switch between strategies on different congruency types or to inhibit

whole number knowledge beyond magnitude information, then we would expect inhibitory control to contribute to both congruent and incongruent pairs. In line with this second possibility, some children may learn that fractions operate “differently” than whole numbers (Miller Singley et al., 2020; Rinne et al., 2017) so they may monitor responses more carefully, therefore engaging inhibitory control on both congruency types.

Additionally, it is possible that the task used to measure inhibitory control may affect the relations to rational numbers that are found. Although the majority of studies examining inhibitory control have found positive associations with rational number outcomes (Abreu-Mendoza et al., 2020; Avgerinou & Tolmie, 2019; Coulanges et al., 2021; D. M. Gómez et al., 2015; Ren & Gunderson, 2021), some studies have not found significant relations between the measures (Matthews et al., 2016; Park & Matthews, 2021; Stricker et al., 2021). Notably, Matthews and colleagues found that inhibitory control was not related to performance on a non-symbolic fraction comparison task or to conceptual understanding of rational numbers, nor was it related to symbolic fraction comparison (Matthews et al., 2016; Park & Matthews, 2021). These conflicting results may stem from the type of inhibition task employed. While these studies (Matthews et al., 2016; Park & Matthews, 2021) used an arrow Flanker task, studies that demonstrated relations with rational outcomes used variations of the Stroop task (Avgerinou & Tolmie, 2019; Coulanges et al., 2021; D. M. Gómez et al., 2015) or the Hearts and Flowers task (Abreu-Mendoza et al., 2020; Ren & Gunderson, 2021). One possible explanation for these varying results is that inhibitory control may be a diverse construct of which common inhibitory control tasks capture different subcomponents of this capacity. A prominent division of inhibitory constructs actually considers both Stroop and Flanker as measures of the same subcomponent, variously called “response-distractor inhibition” (distinct from resistance to proactive interference) (Friedman & Miyake, 2004) or “interference resolution” (Younger et al., 2023). Based on this grouping, the disparate findings for Stroop and Flanker as they relate to rational number understanding are difficult to parse. However, within the rational number field a competing framework contrasts semantic inhibition, (i.e., overcoming learned knowledge) from response inhibition (i.e., overcoming prepotent responses) (Avgerinou & Tolmie, 2019). In this view, Stroop is a measure of semantic interference, whereas Flanker can be seen as a response inhibition measure, specifically one that involves overcoming visual distractors (K. Lee & Lee, 2019). Based on this division, stronger prior findings for Stroop compared to Flanker suggest that semantic interference may be the key capacity tapped by rational numbers. These conflicting organizing schemes suggest that additional work is needed to better understand how inhibitory control contributes to the development of fraction understanding.

In the current study, we selected two domain-general inhibitory control tasks, color-word Stroop task and letter Flanker. Given the diversity of measures that have been examined in relation to rational number comparison skills and the lack of clarity on which measures should have the largest contributions, we created a composite measure of these two tasks to capture individual differences in inhibitory control that are not specific to any one task. In a follow-up, supplementary analysis, we examined which measure had the stronger contribution to fraction comparison performance.

Whole number knowledge: help or hindrance?

A large body of evidence relates better symbolic whole number knowledge to higher mathematical achievement (Schneider et al., 2017; Smedt et al., 2013). However, the whole number interference hypothesis suggests that in the case of fractions, better symbolic whole number knowledge may actually impede success with fraction comparison. Fractions have a bipartite structure (a/b), in which the numerator and denominator enter into a multiplicative relation. This structure affords at least two different types of processing during fraction comparison: 1) holistic processing, comparing the magnitudes of the two fractions, and 2) componential processing, comparing the whole number components of the two fractions (Ischebeck et al., 2009; Meert et al., 2009; Miller Singley & Bunge, 2018; L. Zhang et al., 2014). Componential processing could create a context in which whole number knowledge bolsters performance on congruent comparisons but interferes with performance on incongruent comparisons.

Whole number knowledge is typically measured using two different types of tasks: number line estimation, in which participants estimate where a number goes on a number line, and number comparison, in which participants make a speeded judgment as to which of two numbers has the larger magnitude. Although performance on these two measures is typically correlated (Laski & Siegler, 2007; Ramani & Siegler, 2008), they may tap into separable aspects of whole number magnitude knowledge. Potentially, whole number line estimation may index a learner's capacity to linearly organize numerical magnitudes (Siegler et al., 2011). By contrast, whole number comparison may index the automaticity of accessing magnitude from numerical symbols. Because properties that apply to whole numbers do not always apply to fractions, fractions may represent a counterintuitive type of math knowledge for which this whole number automaticity, which is typically beneficial, may in fact be a hindrance (Bonato et al., 2007; Ischebeck et al., 2009; Rosenberg-Lee, 2021; Vamvakoussi & Vosniadou, 2004; L. Zhang et al., 2014). Specifically, whole number magnitude comparison may contribute to componential processing during fraction comparison and thus represent a step where interference from whole numbers may come into play. For this reason, in the current study we are particularly interested in single-digit whole number comparison. We examined whether this measure predicts fraction comparison performance for congruent and incongruent problems with both shared and distinct component pairs. If whole number comparison taps the automaticity of whole number magnitude understanding, we might expect it to have a positive contribution to congruent problems but a weaker or negative contribution for incongruent problems because a focus on the magnitude of fraction components in these problems will lead to the incorrect answer. In contrast, if whole number comparison measures a deep understanding of numerical magnitude, regardless of number system (i.e., whole or rational), we would expect a positive contribution across congruency types.

Further, we examined whether these contributions differ by fraction pair component type (shared vs. distinct). On the one hand, opportunity for whole number interference is greater for distinct component problems (which require comparing multiple components). On the other hand, the presence of incongruent information may be more salient for the shared component problems. We explicitly considered whether whole number knowledge influences fraction performance

over and above the contributions of working memory and inhibitory control. This approach allows us to begin separating out the need for inhibition from the automatic activation of the information to be inhibited.

Tracking the development of fraction understanding

In the United States, fractions are typically introduced in 3rd grade, and students should have developed considerable proficiency by the end of middle school in 8th grade (Common Core State Standards Initiative, 2010). Although several studies have considered the development of fraction knowledge over this time period (Kainulainen et al., 2017; Van Hoof et al., 2018), none to our knowledge have considered the changing role of cognitive building blocks like EFs or whole number knowledge, which continue to develop and improve over this age range (Constantinidis & Luna, 2019; Opfer & Siegler, 2007). As EFs improve, they may play a bigger role in supporting fraction performance. For example, stronger EFs may be needed more in later grades when students are implementing more sophisticated strategies. Alternatively, EFs may be more important in earlier grades when students are first learning fraction concepts. The interplay between these EFs and whole number knowledge may cause the interference effects on fraction comparison to grow as well. To investigate these relations, we analyzed cross-sectional data from 3rd, 5th, and 7th graders that were collected as part of a larger, longitudinal study (Younger et al., 2022, 2023). This cross-sectional sample affords a view across the first five years of fraction instruction in the US and allows us to investigate how EFs and whole number knowledge relate to fraction comparison at three different stages in the development of these skills.

The current study

The primary aim of this study was to test two *a priori* predictions of the whole number interference hypothesis. First, if inhibition is vital for rational number understanding, we would expect inhibitory control to be an independent predictor of fraction comparison performance after accounting for working memory. Second, if difficulties with rational numbers stem from whole number magnitude interference, then we would expect whole number comparison to be negatively associated with fraction comparison performance. For both predictions, we explored how fraction pair congruency type (congruent or incongruent with whole number knowledge) influenced the relations between the cognitive factors. We also examined whether inhibitory control and whole number knowledge interacted to determine whether participants with poorer whole number knowledge require less inhibition. Finally, we explored whether the relations between inhibitory control, whole number knowledge, and fraction comparison were stable across grades or changed over development. Together, these analyses paint a comprehensive picture of the role of inhibitory control and whole number knowledge during the crucial early years of the development of fraction understanding.

Methods

Project iLead Study

A total of 1,280 3rd through 8th grade children participated in the Project iLead study, which was a longitudinal study over two school years (2016-2017 and 2017-2018) that investigated EF development in elementary and middle school (Younger et al., 2022, 2023). Participants were recruited from nine schools in northern California—seven public schools in one district (5

elementary K-5; 2 middle 6-8), one parochial K-8 school, and one private K-8 school. Of the seven public schools, two of the elementary schools and one of the middle schools were Title 1 schools, a designation under US law indicating that these schools had high percentages of students from low-income families and received federal funds to support these students.

Participants were assessed in the fall and spring of each school year, for a total of four assessment periods. Each assessment period consisted of two sessions, one for the EF assessments and a second for an academic performance assessment. This current study focuses only on year 1 of the study, when students from grades 3, 5, and 7 were recruited (see Younger et al., 2022 for more details about study enrollment).

Participants

The data described here are from 765 students who provided sufficient data for our tasks of interest in the first year of the project. The sample was ethnically and socioeconomically diverse (Table 2.2.1). The study was performed in accordance with the Institutional Review Board at the

Table 2.2.1. Demographic characteristics of the sample. See Supplementary Methods for details on the sample size difference between grades. See Table 2.2.S1 for age at testing separated by timepoint.

Variable	Grade 3 n=180	Grade 5 n=163	Grade 7 n=422
Age at testing (years)	8.80 (7.99, 10.34)	10.61 (9.83, 11.72)	12.63 (11.30, 14.59)
Not reported	4.4%	3.7%	0.5%
Gender			
F	47.8%	35.0%	46.0%
M	38.9%	44.8%	47.4%
Not reported	13.3%	20.2%	6.6%
Ethnicity			
American Indian or Alaskan Native	0.0%	0.0%	1.2%
Asian	38.3%	26.4%	31.0%
Black or African American	2.8%	3.7%	1.2%
Blank on Purpose	0.0%	0.0%	0.2%
Filipino	3.3%	6.1%	7.8%
Hispanic or Latino	18.3%	23.9%	31.0%
Pacific Islander	1.7%	0.0%	0.7%
Two or More Races	5.6%	4.3%	4.3%
White	16.7%	15.3%	15.9%
Not reported	13.3%	20.2%	6.6%
Free/Reduced Lunch	27.2%	33.7%	33.9%
Not reported	13.3%	20.2%	6.6%
Mean (Range); %			

University of California, San Francisco. Written parental or guardian consent was obtained from all participants at the beginning of the study, and verbal assent from all participants was obtained before all in class data collection sessions. At the end of the study, all students in participating classrooms received snacks and stickers, regardless of their individual participation and performance.

Procedure

Two groups of assessments were administered to participants at two timepoints during each school year, once in the fall and again in the spring. The first group of assessments was the Adaptive Cognitive Evaluation (ACE) battery (Younger et al., 2021), which consisted of nine EF tasks and was administered at each timepoint. The second group were scholastic assessments and were divided into two subsets (seven to eight tasks each), one of which included the numerical comparison tasks. Each subset was administered to participants once per year in alternating semesters. Classrooms were randomly assigned to complete each task set in either the fall or spring of each year, with the other subset of assessments administered at the other timepoint. At each of the timepoints, ACE was administered first, and the scholastic assessments were administered approximately six weeks later ($M=5.7$ weeks, $min.=1.9$, $max.=10$). All tasks were administered on iPads in a group setting ($M=30$ students, $min.=7$, $max.=83$) at schools during approximately 50-minute sessions. Tasks were designed to be as short as possible given the number of assessments that needed to fit into each of the sessions. The number of facilitating researchers ranged from four to 12, depending on group size, and administration took place in various school contexts, including classrooms, libraries, cafeterias, and gymnasiums. Instructions were given verbally to the whole group by the lead facilitator, with complementary visual instructions on a 24" x 36" color flipbook in front of the class. All participants in a group began each task at the same time. When every participant had completed the task, the lead facilitator provided instructions for the next task. Before completing each task, participants completed a few practice trials and could ask the researchers questions. Researchers monitored the session throughout.

Executive function tasks

Working memory tasks. The forward spatial span task was adapted from the Corsi Block Task (Corsi, 1972) for use on touch-screen tablets. Participants had to reproduce a spatial sequence of illuminated locations. The backward spatial span task had the same design as forward spatial span, but participants were prompted to recall the sequence in the reverse order. For both tasks, participants started at a span of 3, which is a common starting point for Corsi Block Tasks for children in grades 3 and above (Farrell Pagulayan et al., 2006). The longest attempted span was used as the outcome measure. For more details on these tasks, see Supplementary Methods and Figure 2.2.S1.

Inhibitory control tasks. The color-word Stroop task was a modified computerized version of the Stroop paradigm (Stroop, 1935) for manual responses (Mead et al., 2002). It consisted of 50 experimental trials, 70% congruent (font color matches color word) and 30% incongruent (font color differs from color word). The letter Flanker task, which used the letters "A" through "D", was a computerized adaption of the original Eriksen and Eriksen (1974) task that also used these

letters. The task consisted of 50 experimental trials, of which 50% were incongruent and 50% were congruent. Response windows for both tasks were adaptive to keep them challenging for all ages of participants (see Younger et al., 2022 for more on adaptivity). In part due to this adaptivity, the outcome measure used was rate correct score (RCS; Vandierendonck, 2017; Woltz & Was, 2006) across the full task: the number of correct responses per unit time, calculated by dividing the total number of correct responses by the total response time in seconds summed across all trials of the task, both congruent and incongruent. This measure accounts for speed-accuracy tradeoffs and has been used in prior work on inhibitory control (Coulanges et al., 2021). For additional details on the tasks and choice of metric, see Supplementary Methods and Figure 2.2.S2.

Numerical comparison tasks

For the whole number and fraction comparison tasks, participants always saw two numbers, one on the left side of the screen and one on the right, and were instructed to tap the larger of the two numbers. Within each task and level, the position of the larger number (i.e., on the left or right) was counterbalanced and trials were presented in a random order. Trials that were not answered within the response window were marked as incorrect. Whole number comparison was administered first, followed by fraction comparison.

Whole number comparison. In this task, the stimuli were two single-digit whole numbers, and all pairs had a magnitude difference of 1 (see Figure 2.2.S3). Thus, the task had eight different comparison problems—1 versus 2, 2 versus 3, and so on, up to 8 versus 9. Each problem was presented four times, twice with the larger number on the left and twice with the larger number on the right, yielding 32 trials. However, due to a programming error, participants only saw 30 trials. At the fall timepoint, participants were missing the two presentations of 7 versus 8 with the larger digit on the right. At the spring timepoint, participants were missing two trials with the larger digit on the right, one presentation of 6 versus 7 and one of 7 versus 8. The response window was three seconds.

Fraction comparison. We designed this task to have two levels of difficulty (Figure 2.2.1) because this study included a broad range of grades and therefore a broad range of expected fraction knowledge. Level 1 consisted of 16 unique shared component trials: eight congruent, same denominator problems (e.g., $4/9$ vs $7/9$) and eight incongruent, same numerator problems (e.g., $3/4$ vs $3/7$). In half of the trials for each congruency type, the larger number was presented on the left. In selecting fraction pairs for the congruency types, we controlled for the average magnitude distance between the two fractions and the partial distance (either 1 or 3) between the digits in the non shared component. Further, fractions were limited to those that could not be reduced to lower terms (see Table 2.2.S2 for a full list of stimuli and additional properties).

To prevent students from feeling discouraged by seeing problems beyond their ability level (e.g., material they were not yet expected to know based on the national math standards), only participants who achieved at least 75% accuracy (i.e., ≥ 12 problems correct) moved on to Level 2. Participants who did not meet this mark completed Level 1 again, though we only analyzed data from their first time through the task. 75% accuracy on Level 1 was chosen as the criterion

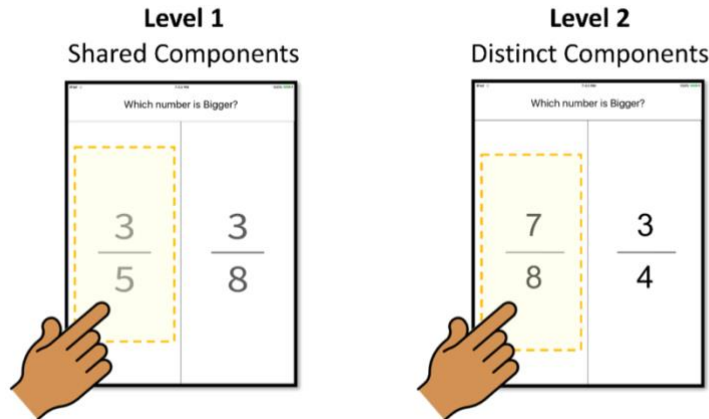


Figure 2.2.1. Two sample trials from the fraction comparison task. In Level 1 (left), the fraction pair had shared components, either the same denominator or the same numerator. In Level 2 (right), the fraction pair had distinct components. If participants reached 75% accuracy on Level 1, they advanced to Level 2. The response window for both levels was 4 seconds. The hands and yellow boxes are for illustrative purposes only and indicate the correct response option for each trial.

in order to ensure that participants who advanced had some mastery of Level 1 (i.e., reliably more accurate than chance at 50%) without excluding too many participants (i.e., including only those at ceiling) in order to ensure sufficient variability on Level 2.

Level 2 consisted of 10 unique problems with distinct components (e.g., $\frac{3}{4}$ vs $\frac{5}{6}$). Each problem was presented twice to counterbalance which fraction in the pair was presented on the left, yielding 20 trials. Half of the trials were congruent with whole number knowledge, such that the fraction with the larger numerator and denominator had the larger magnitude (e.g., $\frac{3}{7}$ vs $\frac{4}{8}$). The other half were incongruent with whole number knowledge, such that the fraction with the larger numerator and denominator had the smaller magnitude (e.g., $\frac{2}{3}$ vs $\frac{5}{9}$). In selecting fraction pairs for the congruency types, we again controlled for the magnitude distance between the two fractions. We were also able to control for the partial distance for the numerators (which ranged from 1-3 for both congruent and incongruent), but mathematically (Rosenberg-Lee, 2021) this means we could not control for denominator distance (which ranged from 1-3 for congruent and 3-6 for incongruent pairs) and other features (see Table 2.2.S3 for a full list of stimuli and their properties).

For both levels, we chose a response window of four seconds to encourage automatic processing rather than strategies that involve calculation, such as computing the cross products. Examination of the fraction comparison data suggest that the 4 second response window was sufficient because the average time until a correct response was made was 1.88 seconds for Level 1 and 1.99 seconds for Level 2. Further, only 15% of incorrect trials on either level were due to no responses (i.e., timing out).

Data analysis

Preparation for analysis. First, trial- and task-level data were cleaned based on the procedures outlined in the Supplementary Methods. After following these procedures, we created composite measures for working memory (WM) from performance on the forward and

backward span tasks and for inhibitory control (IC) from performance on the Stroop and Flanker tasks. To compute the composite measures, participants' scores on the EF tasks were first z-scored within grade and timepoint to account for the possibility of scores increasing over the school year. If participants had data for both tasks for each construct (WM: 99.35% of participants; IC: 95.56% of participants), their composite score consisted of the average of the two standardized scores. If participants had data for only one task, their composite score corresponded to the standardized score for that task (WM: 0.52% forward span only and 0.13% backward span only; IC: 1.57% Flanker only and 2.88% Stroop only). Participants with no data for either task were excluded from analysis. The composite measures were then z-scored within each grade and timepoint. For the follow-up analyses of the independent contributions of Stroop and Flanker, only participants with data for both tasks were included in the analysis (see Supplementary Methods).

Only participants who scored above chance (50%) on whole number comparison were included in analyses (24 participants excluded), as we reasoned that participants who scored below chance likely either did not understand the instructions or chose not to follow them. Whole number comparison accuracy was z-scored within grade and timepoint, to account for differences in the stimuli and so that all predictors were standardized, for subsequent modelling.

Next, because we used fraction comparison congruency type (congruent or incongruent) as a fixed effect in the mixed effects models, we needed to ensure that participants had enough trials of each type to analyze. Therefore, we included in our analyses only participants who had data for at least 75% of the trials for each congruency type after trial-level data cleaning (Fraction Comparison Level 1: ≥ 6 trials of each type, excluded 13 participants; Fraction Comparison Level 2: ≥ 8 trials of each type, excluded 18 participants).

Finally, participants who had data for all of our primary tasks of interest (Fraction Comparison Level 1, WM, IC, and Whole Number Comparison) were included in the final sample of 765 participants (Table 2.2.2). Given that at least 75% accuracy on Fractions Level 1 was required to advance to Level 2, a subset of this final sample had data for Level 2¹. The sample with Level 2 data consisted of 473 participants (Table 2.2.2).

Analysis methods. We applied mixed effect models to examine the relationship between the predictor variables and grade with fraction comparison performance. In all models, we implemented grade as an ordered factor (Grade 3, Grade 5, Grade 7), which allowed us to maintain the rank order of increasing grade levels while also acknowledging that grade cannot be treated as a continuous variable. Therefore, the regression models used a polynomial contrast,

¹ Due to the need for participants to have enough trials of each congruency type to be included in the analyses for each level, a small number of participants (N=6) scored above 75% on Level 1 and completed Level 2 but were excluded from the Level 1 analyses for insufficient trials of each congruency type. These participants were still included in the Level 2 analyses because they did have the sufficient number of trials for each congruency type at that level. The results did not change if the 6 participants were excluded from the Level 2 analyses. See the Supplementary Methods for additional details about how many trials for each congruency type were considered sufficient.

Table 2.2.2. Sample sizes by grade level and fraction comparison component type. In order to complete fraction comparison with distinct components (Level 2), participants needed to achieve 75% accuracy on fraction comparison with shared components (Level 1). The percentage of participants within each grade that advanced to Level 2 is shown in parentheses.

Fraction Comparison	Grade 3	Grade 5	Grade 7	Total
Shared components (Level 1)	180	163	422	765
Distinct components (Level 2)	55 (31%)	98 (60%)	320 (76%)	473 (62%)

which outputs linear and quadratic effects for grade and interactions with grade. Implementing grade as an ordered factor allowed us to capture differences in performance across grades that may be better characterized by a quadratic compared to a linear function, for example in the case of floor effects in lower grades or ceiling effects in higher graders.

For our main analyses, we employed logistic mixed effects models predicting fraction comparison trial-level accuracy. All mixed effects models had the same fixed and random effects structure: fixed effects of grade level (Grade 3, Grade 5, Grade 7), congruency type (congruent or incongruent), and their interaction, plus a random intercept for participant and a random slope for congruency type. We will refer to this model as the base model.

To this base model, we first added the WM composite score and then added the IC composite score. Model comparison between the WM model and the WM + IC model addressed the question of whether IC is an independent predictor of fraction comparison accuracy, explaining additional variance over and above WM. We then added whole number comparison and compared this model to the WM + IC model to understand the relation between whole number comparison and fraction comparison after accounting for WM and IC. Successive models were compared with likelihood ratio tests using the anova function (`lmerTest`). To explore possible interactions between the predictors of interest, we also constructed a model of fraction comparison accuracy with interactions between grade, congruency type, IC, and whole number comparison. In cases where this model indicated significant interactions, we computed follow-up analyses to better understand the sources of interactions. Finally, we ran follow-up analyses to examine Stroop and Flanker as independent predictors of fraction accuracy (see Supplementary Methods).

Because we used logistic mixed effects regressions predicting accuracy, the resulting linear models specified the log odds of getting a trial correct. However, it is more interpretable to describe the results in terms of the odds by exponentiating the beta coefficients (e^{β_x}), rather than keeping it in terms of the log odds. For slope coefficients, e^{β_x} gives the odds ratio, relative to 1. When $e^{\beta_x} > 1$, $(e^{\beta_x} - 1) \times 100$ gives the percent increase in the odds of getting a trial correct

for every 1-unit increase in the predictor. When $0 < e^{\beta x} < 1$, $(1 - e^{\beta x}) \times 100$ represents the percent decrease in the odds for every 1-unit increase in the predictor.

Transparency and openness

This study's design and its analysis were not pre-registered. The data analyzed in this study represent a subset of the data collected, and data from the other measures will be reported elsewhere (Younger et al., 2022, 2023). We report all data exclusions and all manipulations for the tasks of interest. The tasks and data are available from Project iLead by request (<https://sites.google.com/view/projectileadnsf>), and the data will be made publicly available by Project iLead two years after publication. Data cleaning and analysis scripts are available on OSF (<https://osf.io/z7hxa/>). All analyses were conducted in R Version 4.2.1 (R Core Team, 2021). We used the *lmerTest* package to conduct the mixed effects models (Kuznetsova et al., 2017) and *emmeans* Version 1.8.2 to conduct post hoc Tukey comparisons (Lenth, 2022). We used *ggplot* Version 3.3.6 (Wickham, 2016), *ggbeeswarm* Version 0.6.0 (E. Clarke & Sherrill-Mix, 2017), see Version 0.7.3 (Lüdtke et al., 2021), and source code from *raincloudplots* (Allen et al., 2021) to visualize the data.

Results

Effect of grade on predictor and outcome measures

Executive functions and whole number comparison. Performance on each of the EF predictor measures improved with grade (Figure 2.2.2A-D; see Supplementary Results for details of the linear regression analyses). Preliminary analyses of accuracy and response time on Stroop

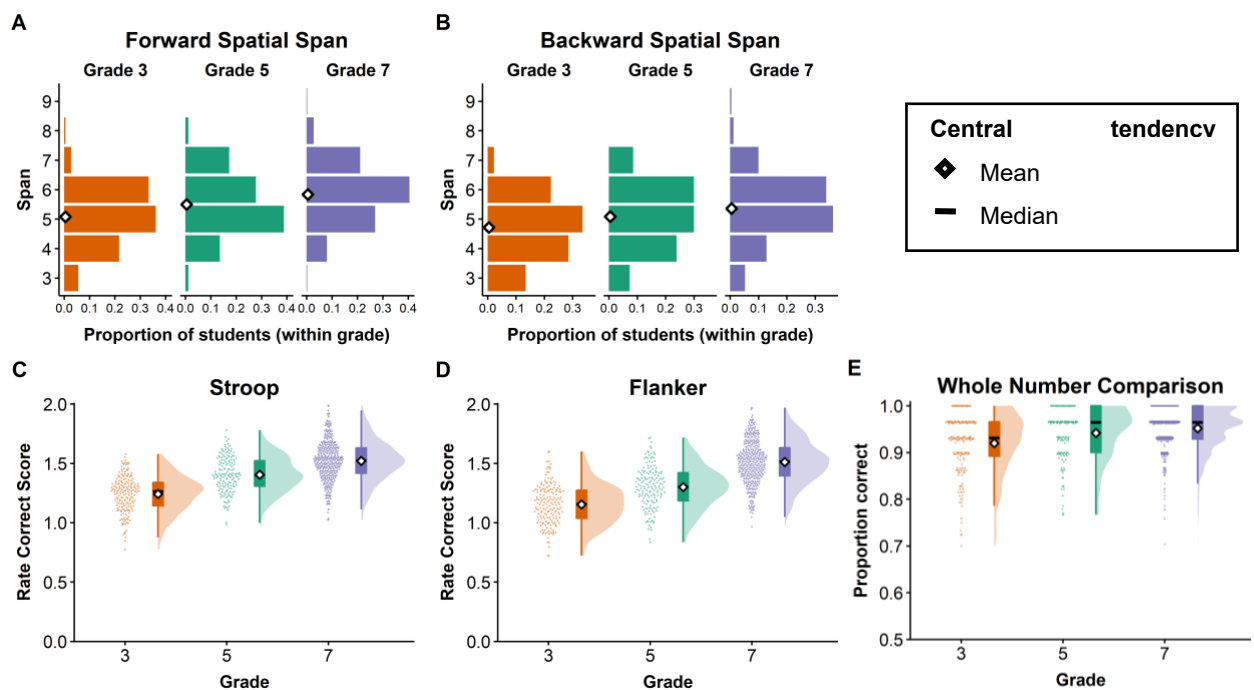


Figure 2.2.2. Distribution of performance by grade for the two working memory tasks, (A) forward and (B) backward spatial span, the two inhibitory control tasks (C) color-word Stroop and (D) letter Flanker, and (E) whole number comparison. Diamonds indicate mean and black horizontal lines indicate medians. If the mean and median values are very close together, the diamond occludes the horizontal line.

and Flanker showed the expected congruency effects (Table 2.2.S4; Figure 2.2.S4), and RCS was used as the outcome measure for these tasks for the rest of the analyses. For whole number comparison, accuracy ranged from an average of 92% in 3rd grade to 95% in 7th grade. Despite this limited range, a linear model predicting accuracy on this task revealed a significant linear effect of grade ($B=0.02$, $SE=0.003$, $p<.001$), which indicates that accuracy on whole number comparison increased linearly with grade, with no quadratic effect ($B=-0.005$, $SE=0.004$, $p=.242$), see Figure 2.2.2E.

Fraction comparison with shared components (Level 1). For fraction comparison with shared components (Level 1), congruent problems have the same denominator and incongruent problems have the same numerator. On average, participants in all three grades performed above chance (50%) on both congruency types ($ts>3.90$, $ps<.001$; see Figure 2.2.3). To test the effects of grade on accuracy, we employed the logistic mixed effects base model described previously: fixed effects of grade level (Grade 3, Grade 5, Grade 7), congruency type (congruent or incongruent), and their interaction, with a random intercept for participant and a random slope for congruency type. This model revealed a conditional effect of grade ($B=1.11$, $SE=0.11$, $p<.001$), which reflects the expected increase in accuracy with grade, and a conditional effect of congruency type ($B=-0.74$, $SE=0.08$, $p<.001$), which reflects the expected lower accuracy for same numerator problems compared to same denominator problems. There was also a significant quadratic effect of grade ($B=-0.35$, $SE=0.13$, $p=.007$), which significantly interacted with congruency type ($B=0.41$, $SE=0.14$, $p=.003$), reflecting plateauing accuracy for same denominator problems in grade 5 versus a linear increase in accuracy for same numerator problems across grades 3, 5, and 7 (see Figure 2.2.S5 for an illustration of the grade effects).

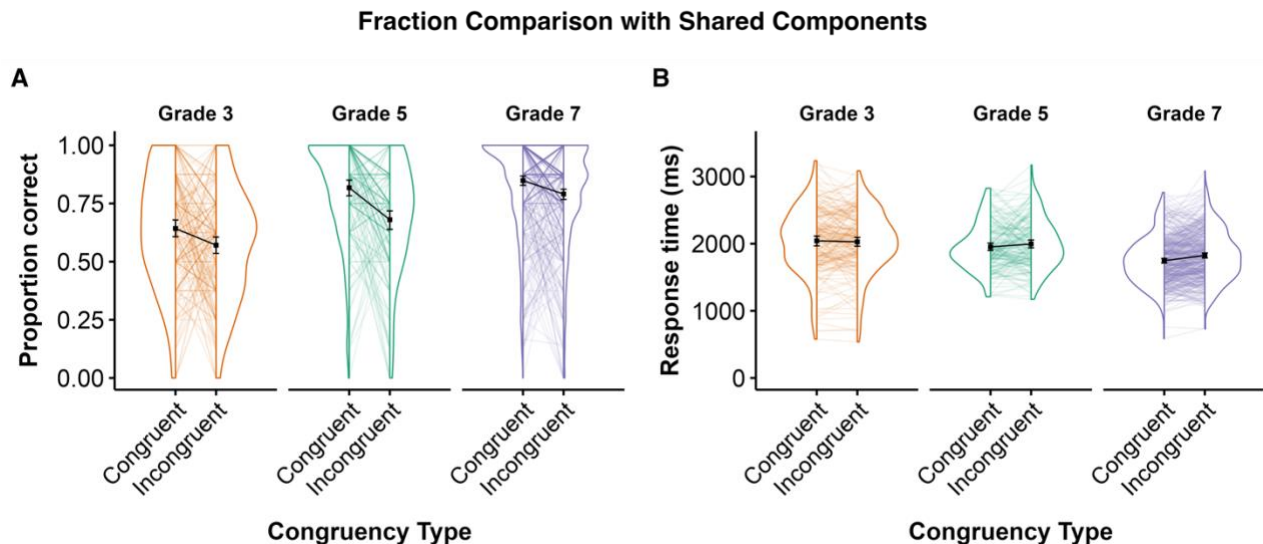


Figure 2.2.3. (A) Accuracy and (B) response time distributions for fraction comparison with shared components (Level 1). Congruent problems have shared denominators and incongruent problems have shared numerators. Each colored line connects a participant's averages on congruent and incongruent problems. The black points represent the overall mean, and the black lines connect these means. Error bars indicate the bootstrapped 95% confidence intervals around the means.

To examine the effect of grade on RTs, we employed a linear mixed effects model with the same fixed and random effects structure predicting RTs on correct trials. This model revealed a conditional effect of grade ($B=-220.18$, $SE=24.88$, $p<.001$), which reflects the expected linear decrease in RTs with increasing grade. Overall, participants were slower on same numerator problems than on same denominator problems ($B=75.23$, $SE=14.22$, $p<.001$, Figure 2.2.3), and there was a significant interaction between the linear grade effect and congruency type ($B=68.41$, $SE=23.17$, $p=.003$, see Figure 2.2.5). Post-hoc grade-wise Tukey comparisons showed that this effect was driven by slower RTs for same numerator relative to same denominator trials in grades 5 (*Estimated Marginal Mean (EMM)*=-97.91, $SE=27.32$, $p=.005$) and 7 (*EMM*=-112.27, $SE=16.19$, $p<.001$), but no difference in grade 3 (*EMM*=-15.52, $SE=28.49$, $p=.994$).

Fraction comparison with distinct components (Level 2). Contrary to our expectation, participants performed better on incongruent compared to congruent fraction comparison problems with distinct components (Level 2). A logistic mixed effects model predicting accuracy from the previously described base model showed a conditional effect of congruency type ($B=1.18$, $SE=0.11$, $p<.001$; see Figure 2.2.4), confirming that performance on incongruent problems was more accurate than on congruent problems, and there were no conditional effects of grade (B s=0.11 and 0.09 for linear and quadratic effects, respectively, p s>0.366). Further, there was a significant interaction between congruency type and linear effect of grade ($B=0.54$, $SE=0.18$, $p=.003$). Given that congruent was the reference category, the lack of conditional effect of grade but presence of an interaction indicates no improvement across grade for congruent problems and a linear improvement across grade for incongruent problems (see Figure 2.2.5 for clearer illustration of grade effects).

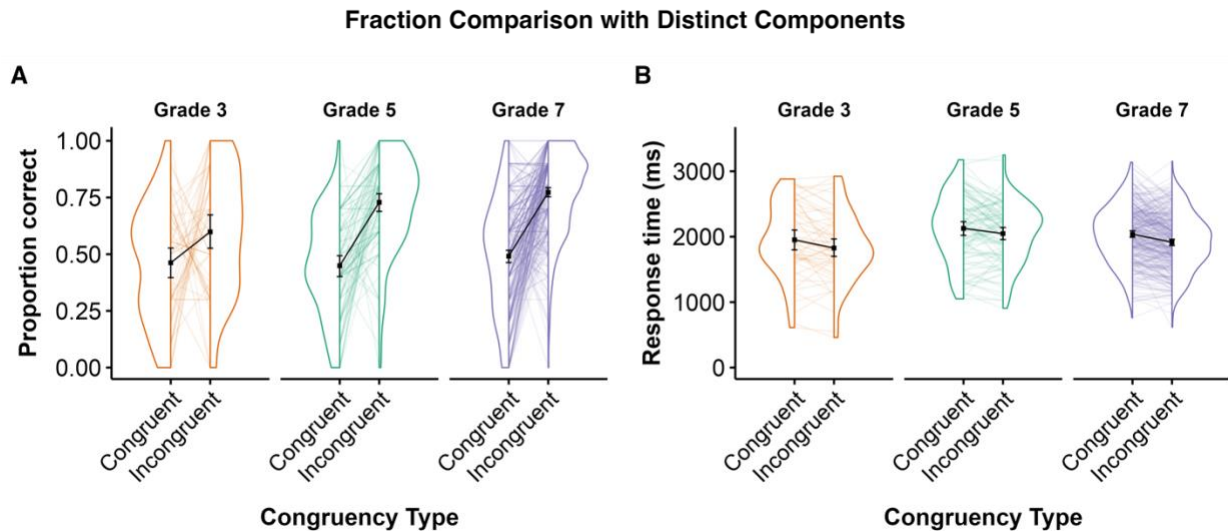


Figure 2.2.4. (A) Accuracy and (B) response time distributions for fraction comparison with distinct components (Level 2). Each colored line connects a participant's averages on congruent and incongruent problems. The black points represent the overall mean, and the black lines connect these means. Error bars indicate the bootstrapped 95% confidence intervals around the means.

Table 2.2.3. Beta coefficients for logistic mixed effects models predicting accuracy on fraction comparison with shared components (Level 1) from our predictors of interest. The base model includes fixed effects of grade level (3, 5, 7), congruency type (congruent or incongruent), and their interaction, plus a random intercept for participant and a random slope for congruency type. WM=Working Memory Composite Score, IC=Inhibitory Control Composite Score, and WNC=Whole Number Comparison. Congruent problems had the same denominator and incongruent problems had the same numerator. (See Table 2.2.S5 for 95% confidence intervals)

Predictor	Base	Base + WM	Base + WM + IC	Base + WM + IC + WNC
Grade (linear)	1.11***	1.10***	1.10***	1.12***
Grade (quadratic)	-0.35**	-0.35**	-0.35**	-0.36**
Congruency type	-0.74***	-0.74***	-0.74***	-0.75***
Grade (linear) x Congruency type	-0.21	-0.20	-0.20	-0.21
Grade (quadratic) x Congruency type	0.41**	0.41**	0.41**	0.41**
Working memory		0.29***	0.16***	0.13**
Inhibitory control			0.34***	0.30***
Whole number comparison				0.27***
AIC	11778.50	11735.67	11683.07	11644.62
BIC	11844.91	11809.45	11764.23	11733.16

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

A linear mixed effects model predicting RTs on correct Level 2 trials revealed that participants were also faster on incongruent problems than on congruent problems ($B = -111.21$, $SE = 26.02$, $p < .001$, Figure 2.2.4). There was also a significant quadratic effect of grade ($B = -136.16$, $SE = 53.27$, $p = .011$), capturing the inverted-U shaped pattern of increasing RTs from grades 3 to 5 and decreasing RTs from grades 5 to 7 (Figure 2.2.S6).

Predicting performance on fraction comparison with shared components (Level 1)

In the next series of analyses, we investigated how the three predictors of interest—WM, IC, and whole number comparison—related to accuracy on fraction comparison with shared components (Level 1). Given the significant correlations between most of our variables of interest (Figure 2.2.5A), we employed a series of logistic mixed effects models to directly address our primary research questions: namely, the independent contributions of IC and whole number knowledge on fraction comparison ability. Building on the base model, we successively added the WM composite score, the IC composite score, and standardized whole number comparison accuracy.

The model including WM explained the data better than the base model ($\chi^2(1) = 44.83$, $p < .001$). Every increase in WM score of one standard deviation was associated with a 34% increase in the odds of getting a trial correct (Table 2.2.3). The model adding IC better explained the data than the model with only WM as a covariate ($\chi^2(1) = 54.60$, $p < .001$), indicating that IC was an independent predictor of fraction comparison accuracy, over and above WM. Every increase in IC score of one standard deviation was associated with a 41% increase in the odds of getting a trial correct (Table 2.2.3, Figure 2.2.5B). With the inclusion of IC, WM remained a significant predictor, indicating that the two types of EF measures make independent contributions to fraction performance. The model that additionally included whole number comparison explained the data

Predictors of Fraction Comparison Performance Shared Components (Level 1)

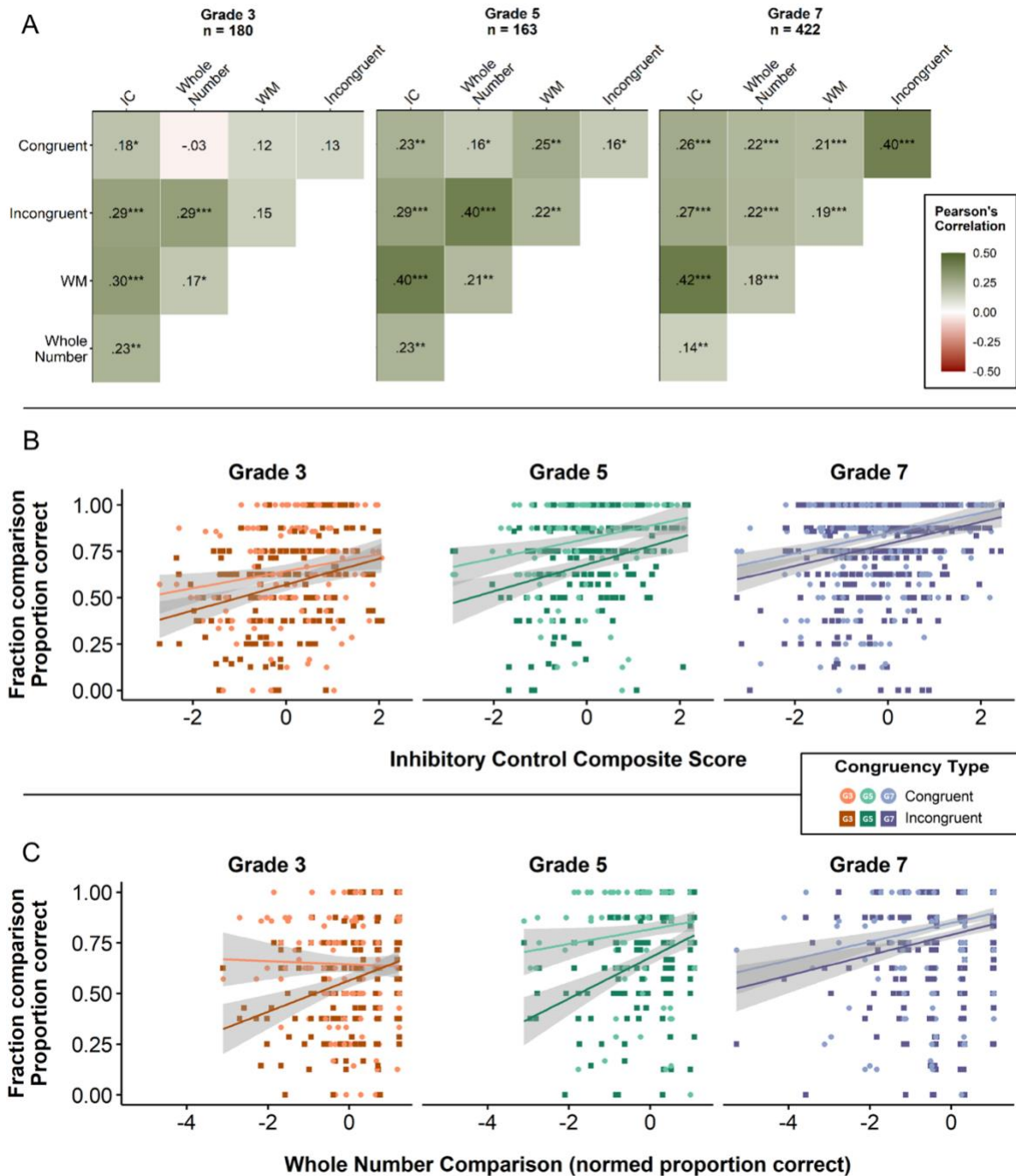


Figure 2.2.5. (A) Zero-order Pearson correlation coefficients between fraction comparison with shared components (Level 1) accuracy on each congruency type and the predictors of interest: working memory composite score (WM), inhibitory control composite score (IC), and whole number comparison accuracy. Uncorrected p-values: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ (B) Relation between inhibitory control composite score and accuracy for congruent and incongruent trials. (C) Relation between whole number comparison and accuracy for congruent and incongruent trials. For (B) and (C), points and lines are shaded by congruency type, with the lighter color points and lines representing congruent (same denominator) problems and the darker color points and lines representing incongruent (same numerator) problems.

better than the model with only WM and IC as covariates ($\chi^2(1)=40.45, p<.001$), showing that whole number knowledge predicts fraction accuracy after taking WM and IC into account. Every increase in whole number comparison accuracy of one standard deviation was associated with a 31% increase in the odds of getting a trial correct (Table 2.2.3, Figure 2.2.5C). This series of analyses suggested that working memory, inhibitory control, and whole number comparison are all unique predictors of fraction comparison accuracy.

To determine if there were any interactions between grade, congruency type, IC, and whole number comparison, we constructed a model with the four-way interaction and the corresponding three- and two-way interactions (see Table 2.2.S6). WM was included as a single covariate with no interaction terms. This model revealed a significant three-way interaction between the linear effect of grade, congruency type, and whole number comparison ($B=-0.34, SE=0.12, p=.006$), as well as three two-way interactions (congruency type x whole number comparison: $B=0.18, SE=0.08, p=.018$; linear effect of grade x whole number comparison: $B=0.29, SE=0.11, p=.009$; and quadratic effect of grade x congruency type: $B=0.36, SE=0.14, p=.011$). Notably, none of these interactions included IC.

To examine these interactions with whole number comparison more closely, we implemented separate logistic mixed effects models for each grade. These grade-wise models included congruency type, whole number comparison, and their interaction, as well as WM and IC as covariates, and the same random effects structure described previously (Table 2.2.4). Performance on same numerator problems was significantly worse than on same denominator trials for all grades. For 3rd graders, there was a significant interaction between congruency type and whole number comparison but no conditional effect of whole number comparison; specifically, whole number comparison did not predict accuracy on same denominator problems, but positively predicted accuracy on same numerator problems (Figure 2.2.5A and 2.2.C). For 5th graders, although Figure 2.2.5A shows a positive correlation between whole number comparison and accuracy for both congruency types, it seems that WM and IC accounted for that relation, as there was no conditional effect of whole number comparison nor an interaction with congruency

Table 2.2.4. Beta coefficients for grade-wise logistic mixed effects models predicting accuracy on fraction comparison with shared components (Level 1) from our predictors of interest. (See Table 2.2.S7 for 95% confidence intervals)

Predictor	Grade 3	Grade 5	Grade 7
Congruency type	-0.42***	-1.07***	-0.78***
Working memory	0.07	0.18*	0.13
Inhibitory control	0.24***	0.26**	0.35***
Whole number comparison	-0.11	0.21	0.36***
Congruency type x Whole number comparison	0.39**	0.21	-0.10
AIC	3555.24	2577.36	5498.35
BIC	3608.60	2629.86	5559.41

*** $p<0.001$; ** $p<0.01$; * $p<0.05$

type. Finally, for 7th graders, better whole number comparison predicted better accuracy on both congruency types (Table 2.2.4). These follow up analyses suggest that whole number knowledge may play a different role in fraction comparison across different grades and levels of fraction experience.

Because IC was a robust predictor of fraction performance on Level 1, we conducted a series of follow-up analyses to more closely examine the two measures of IC: Stroop and Flanker (see Supplementary Methods, Figure 2.2.S7A, and Table 2.2.S8). Both Stroop and Flanker individually explained variance in fraction comparison after accounting for WM ($\chi^2=26.61$, $p<.001$ and $\chi^2=37.34$, $p<.001$, respectively). Further, Stroop and Flanker both explained variance in fraction comparison after additionally accounting for the other inhibitory control measure ($\chi^2=10.96$, $p<.001$ and $\chi^2=21.69$, $p<.001$, respectively). These analyses suggest that both measures are independent predictors of fraction comparison accuracy for shared component problems.

Predicting performance on fraction comparison with distinct components (Level 2)

Next, we examined the predictors of performance on fraction comparison with distinct components (Level 2) with a parallel set of analyses to those carried out for Level 1. Only participants who achieved at least 75% accuracy on Level 1 moved onto Level 2. By this criterion, 31% of 3rd graders (55/180), 60% of 5th graders (98/163), and 76% of 7th graders (320/422) advanced to Level 2 (Table 2.2.2). A chi-squared test for trend in proportions indicated that the proportion of participants advancing to Level 2 increased with grade ($\chi^2(1)=107.34$, $p<.001$).

After considering the correlations between our predictor variables and performance on pairs with distinct components (Figure 2.2.6A), we more formally tested for independent contributions of IC and whole number comparison to fraction comparison. Again, we started with the base model and then conducted a series of logistic mix effects models adding additional predictors (Table 2.2.5). The model including WM did not explain the data better than the base model ($\chi^2(1)=2.61$, $p=.106$), indicating that WM was not a significant predictor of fraction accuracy with distinct components. Next, IC was added to the model, and this model explained the data better than the model with only WM as a covariate ($\chi^2(1)=12.83$, $p<.001$), indicating that participants with higher IC scores also performed better on this level. Every increase in IC of one standard deviation was associated with a 15% increase in the odds of getting a trial correct (Table 2.2.5, Figure 2.2.6B). However, adding whole number comparison to the model did not explain the data better than the model with only WM and IC as covariates ($\chi^2(1)=0.08$, $p=.772$), as whole number comparison was not a significant predictor of fraction accuracy at this level (Table 2.2.5, Figure 2.2.6C). This series of analyses suggests that for the more difficult fraction comparison problems with distinct components (Level 2), only individual differences in IC predicted accuracy.

To determine whether there were any interactions between grade, congruency type, IC, and whole number comparison, we again constructed a model with the four-way interaction and the corresponding three- and two-way interactions (see Table 2.2.S10). WM was included as a single covariate with no interaction terms. This model confirmed a significant interaction between linear effect of grade and congruency type ($B=0.61$, $SE=0.23$, $p=.007$), which was already captured in the base model. While visual inspection of Figure 2.2.6B suggests that IC had a larger influence

Predictors of Fraction Comparison Performance Distinct Components (Level 2)

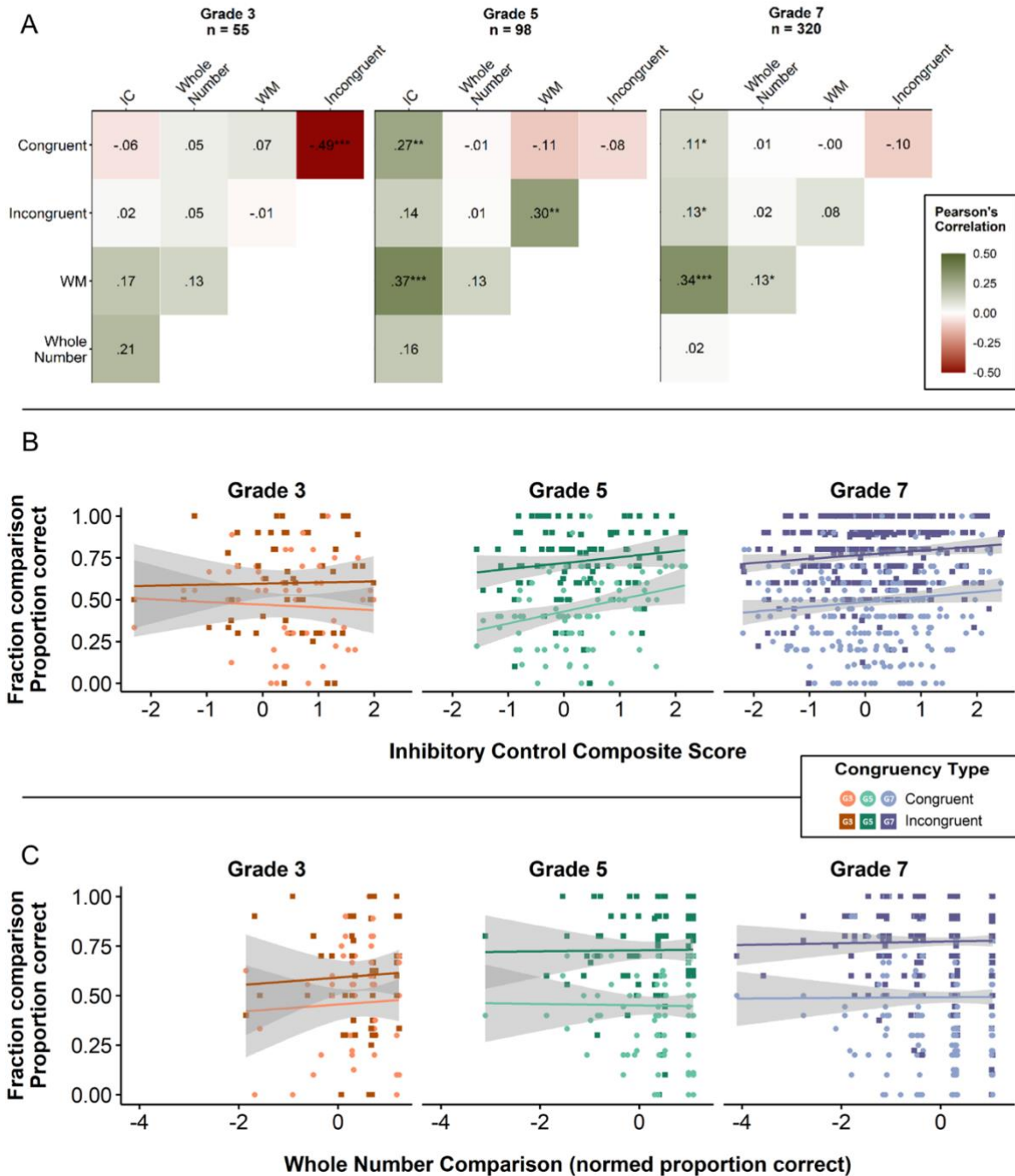


Figure 2.2.6. (A) Zero-order Pearson correlation coefficients between fraction comparison with distinct components (Level 2) accuracy on each congruency type and the predictors of interest: working memory composite score (WM), inhibitory control composite score (IC), and whole number comparison accuracy. Uncorrected p-values: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$ (B) Relation between inhibitory control composite score and accuracy for congruent and incongruent trials. (C) Relation between whole number comparison and accuracy for congruent and incongruent trials. For (B) and (C), points and lines are shaded by congruency type, with the lighter color points and lines representing congruent problems and the darker color points and lines representing incongruent problems.

Table 2.2.5. Beta coefficients for logistic mixed effects models predicting accuracy on fraction comparison with distinct components (Level 2) from our predictors of interest. The base model includes fixed effects of grade level (3, 5, 7), congruency type (congruent or incongruent), and their interaction, plus a random intercept for participant and a random slope for congruency type. WM = Working Memory Composite Score, IC = Inhibitory Control Composite Score, and WNC = Whole Number Comparison. (See Table 2.2.S9 for 95% confidence intervals)

Predictor	Base	Base + WM	Base + WM + IC	Base + WM + IC + WNC
Grade (linear)	0.11	0.11	0.14	0.14
Grade (quadratic)	0.09	0.10	0.09	0.09
Congruency type	1.18***	1.18***	1.18***	1.18***
Grade (linear) x Congruency type	0.54**	0.53**	0.54**	0.54**
Grade (quadratic) x Congruency type	-0.25	-0.25	-0.25	-0.25
Working memory		0.06	0.01	0.01
Inhibitory control			0.14***	0.14***
Whole number comparison				0.01
AIC	11033.33	11032.72	11021.89	11023.81
BIC	11097.45	11103.96	11100.26	11109.30

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

on fraction comparison in grades 5 and 7 than 3, the two-way interaction between quadratic effect of grade and IC did not reach significance ($B=0.13$, $SE=0.14$, $p=.360$). Thus, the results of this interaction model did not warrant grade-wise follow-up analyses.

Because IC was a robust predictor of fraction performance on Level 2, we conducted a series of follow-up analyses to more closely examine the two component measures of IC: Stroop and Flanker (see Supplementary Methods, Figure 2.2.S7B, and Table 2.2.S11). These analyses revealed that both Stroop and Flanker explained variance in fraction comparison after accounting for WM ($\chi^2=10.27$, $p=.001$ and $\chi^2=7.40$, $p=.007$, respectively). Further, Stroop explained variance in fraction comparison with distinct components over and above Flanker ($\chi^2=6.02$, $p=.014$). However, Flanker did not explain variance in fraction performance after accounting for Stroop ($\chi^2=3.14$, $p=.076$). Thus, including Stroop and Flanker as separate predictors revealed an independent contribution to explaining distinct component fraction comparison accuracy for Stroop, but not for Flanker.

Discussion

In this study, we tested two predictions of the whole number interference hypothesis as it relates to fraction comparison in a large, diverse sample of elementary and middle school students. Consistent with the prediction that if part of students' difficulty with mastering fractions comes from the need to inhibit knowledge about whole numbers that does not apply to fractions, we found that individual differences in inhibitory control related to fraction comparison performance. However, we did not find support for the prediction that students with stronger whole number knowledge would experience more interference, and thus, counterintuitively, perform worse on fraction comparison than students with weaker whole number knowledge. These results held for fraction comparisons with both shared and distinct components. Together,

our findings point to a pivotal role for inhibitory control in rational number understanding that may be distinct from inhibiting whole number magnitude knowledge.

Development of fraction comparison performance for 3rd to 7th grade

The youngest participants in our study were in 3rd grade, the grade in which fractions are commonly introduced into the mathematics curriculum in the United States (Common Core State Standards Initiative, 2010). As would be expected based on increasing mastery of fractions with age and accumulating instructional experience, we found that participants became more accurate and responded more quickly on fraction comparison with shared components (Level 1) across grades. As expected, performance on congruent (same denominator) problems exceeded performance on incongruent (same numerator) problems. Further, performance on congruent problems leveled off after 5th grade, whereas performance on incongruent problems increased linearly across grades. However, for the more difficult, distinct component problems (Level 2), contrary to expectation, performance was worse for the congruent than incongruent problems. Further, performance did not improve for congruent problems but did increase linearly across the grades for incongruent problems. Therefore, though there was general improvement across grades, consideration of congruency type revealed subtle performance differences that will be discussed in the following sections.

Inhibitory control supports fraction comparison across grades and problem difficulty

The first aim of this study was to assess the role of inhibitory control in supporting fraction understanding. Although previous studies have documented the contribution of working memory (Jordan et al., 2013) or inhibitory control (Avgerinou & Tolmie, 2019; D. M. Gómez et al., 2015) to fraction comparison performance, none, to our knowledge, have explored the simultaneous contribution of these abilities. We found that when fractions had shared components (Level 1), both working memory and inhibitory control each explained unique variance, and that when fractions had distinct components (Level 2), only inhibitory control explained unique variance. Statistically, this relationship did not interact with grade, although visual inspection suggests it was stronger in grades 5 and 7 than grade 3. A larger sample of 3rd graders may be needed to detect the development of reliance on inhibitory control for distinct component fraction comparison. Nevertheless, these results are in line with a general prediction of the whole number interference hypothesis, namely that inhibitory control should be a vital capacity for rational number understanding.

The strong view of the whole number interference hypothesis, however, would predict that inhibitory control would be especially important for solving fraction comparisons incongruent with whole number knowledge. This prediction was not born out in the present study, as there was an equal contribution of inhibitory control to performance on both congruency types. Prior studies of rational number comparison have found significant contributions of inhibitory control for incongruent but not congruent problems in proportional reasoning (Abreu-Mendoza et al., 2020; Coulanges et al., 2021) and decimal comparison (Coulanges et al., 2021). However, these studies did not explicitly test for the difference in contribution between congruency types, making it unclear whether the contribution of inhibitory control is greater for incongruent than congruent problems. Negative priming studies of fraction comparison also suggest a specific role of

inhibition for incongruent comparisons (Fu et al., 2020; Rossi et al., 2019). These studies show that solving an incongruent, same numerator problem decreases performance on a subsequent congruent, same denominator problem relative to a control prime trial. Notably, these studies involved only shared component problems, making it unclear whether this priming effect extends to distinct component problems where participants might not even know to inhibit their whole number knowledge. Further, they do not consider the effects of congruent priming on incongruent performance, which might be expected if inhibitory control is needed for both problem types. Interestingly, one recent study looked at this question using a block switching design (Van Hoof et al., 2021) and found that switching from congruent to incongruent fractions incurred the same switch cost as the reverse switch, relative to solving the same problem type twice (stay trials). This pattern of results is consistent with our finding that inhibitory control is needed when working with fractions, regardless of congruency type. An interesting area for future work would be to add an inhibitory control assessment to a negative priming study to examine the relation between differences in switch costs and individual variation in this crucial capacity. Another area for future work would be to examine whether the pattern of relations between inhibitory control, congruency type, and performance holds across different types of rational numbers or whether there are fundamental differences between types.

In contrast to the consistent contributions of inhibitory control for both shared and distinct component problems, the role of working memory differed by component type. Working memory had a significant contribution to shared component (Level 1) performance but was not a significant predictor of distinct component performance (Level 2), regardless of whether inhibitory control was present in the model. One explanation is that working memory may support the transition to more automatic processing for easier fraction comparisons (as with shared components), whereas for more challenging tasks (as with distinct components), the ability to inhibit prior knowledge and resolve interference is what matters most. An alternative possibility is that this null effect may stem from the fact that only the strongest performing participants (who included participants with higher working memory) and mostly the oldest participants (5th and 7th grade students who have been learning about fractions for the longest and also have the highest working memory) moved on to Level 2, resulting in a smaller sample size and a possible selection bias. Future studies should use a more fine-grained measure of working memory, such as the half step span measure employed in Coulanges et al., (2021), or a more sensitive task to test working memory's involvement in fraction comparison. Nevertheless, the finding that inhibitory control is a significant predictor of distinct component fraction comparison performance when controlling for working memory further highlights the importance of inhibitory control for fraction understanding. We also found that the contribution of inhibitory control was stable across grades, counter to the possibility that it only becomes important when participants understand that they should be inhibiting their whole number knowledge. This result suggests that even at the earliest stages of fraction understanding, inhibitory control plays a role.

A final question related to inhibitory control afforded by this data set was assessing the contribution of different experimental measures. Prior work has found associations between various Stroop tasks (i.e., numerical and color-word) and rational number outcomes (Avgerinou

& Tolmie, 2019; Coulanges et al., 2021; D. M. Gómez et al., 2015), and also the Hearts and Flowers task (Abreu-Mendoza et al., 2020; Ren & Gunderson, 2021). In fact, the arrow Flanker task is the only measure of inhibitory control studied so far that has not been found to be associated with rational number performance (Matthews et al., 2016; Park & Matthews, 2021). In the current study, both color-word Stroop and letter Flanker were collected. Based on the prior studies and the classification of Stroop as a semantic inhibition task (Avgerinou & Tolmie, 2019), we might have expected that Stroop would have stronger predictive power than Flanker. Consistent with this prediction, for distinct component pairs (Level 2) only Stroop was significant when both were included in the regression analyses, despite both measures predicting performance when entered alone. For shared component pairs (Level 1), both measures predicted performance when entered alone. However, when entered together, the Flanker task had the stronger predictive power, although Stroop remained significant.

Although both color-word (Coulanges et al., 2021) and numerical Stroop (D. M. Gómez et al., 2015) tasks have been previously employed in the literature, here we used the color-word version, which is a domain-independent measure of EF (e.g., it does not involve processing numerical magnitudes), to avoid circularity in interpretations. However, numerical Stroop tasks, which explicitly involve inhibiting whole number magnitude knowledge, may index the automaticity with which numerical magnitude information is accessed (Bugden & Ansari, 2011), making it a potentially useful diagnostic tool to identify students at risk for fraction difficulties. In sum, these results suggest that a wide range of inhibitory control measures can predict rational number outcomes, and more research is warranted to establish if a specific task is more predictive of certain aspects of rational number knowledge and performance, especially in the classroom context.

Whole number knowledge supports some forms of fraction comparison

The second aim of this study was to examine the counterintuitive prediction that better knowledge of whole number magnitudes could be detrimental to fraction understanding. Because properties of whole numbers do not always apply to fractions — for example, the presence of a larger numeral does not always imply the larger fraction magnitude — students who are more able to automatically access whole number magnitudes may experience greater interference when resolving fraction magnitudes, resulting in poor fraction comparison performance. However, this prediction runs contrary to the large body of evidence showing that whole number knowledge is positively related to a variety of math learning outcomes, both when measured concurrently and as a predictor of future learning (Schneider et al., 2017; Smedt et al., 2013).

Consistent with the view that whole number knowledge is beneficial for math achievement, we found that whole number comparison performance was positively related to shared component fraction comparison performance (Level 1). This effect remained significant even after controlling for working memory and inhibitory control, indicating that whole number knowledge contributed over and above domain-general capacities. Interestingly, the effect of whole number knowledge increased with grade level, which may indicate a shift in strategy use. As students gain proficiency

with shared component fractions, they may focus more on comparing the non-shared numbers, such that differences in whole number knowledge become more relevant.

This shift in strategy may help explain the three-way interaction between grade, congruency type, and whole number knowledge that we found for performance on shared component problems (Level 1). Specifically, for 3rd grade, whole number comparison performance predicted fraction comparison on the same numerator problems but not the same denominator problems. There was also a trend in this direction for 5th graders, but by 7th grade, whole number knowledge predicted both congruency types comparably. This result is somewhat unexpected because whole number comparison is most similar to congruent, same denominator comparison, as only numerators must be compared. In contrast, for the incongruent, same numerator problems, the denominators must be compared, and the fraction with the smaller denominator is the larger fraction. Studies that include eye tracking could help elucidate this result by examining whether children with varying whole number knowledge focus differentially on the denominators versus numerators on these types of problems (Miller Singley et al., 2020).

For the more challenging distinct component fraction comparisons (Level 2), we found that whole number knowledge had no significant impact on performance. This lack of relation might reflect the fact that only higher performing students made it to Level 2, and thus there is less variance in the whole number comparison performance in these students (as well as a smaller sample size in which to detect an effect). However, we were still able to detect effects for inhibitory control at this level, suggesting that sampling bias does not preclude finding effects of these factors. Additional work is needed to assess how whole number knowledge relates to various types of fraction knowledge across the ability spectrum. Nevertheless, the lack of a positive contribution of whole number knowledge for fraction comparisons with distinct components suggests that whole number knowledge, as indexed by comparison tasks, may not be equally beneficial for all domains of mathematics.

Most prior work that has demonstrated the utility of whole number knowledge for fractions has employed number line estimation tasks rather than whole number comparison tasks. Indeed, several longitudinal studies employing number line estimation have shown that whole number knowledge is positively related to future performance on a variety of broad measures of rational number understanding, such as conceptual and procedural knowledge of fraction arithmetic (Bailey et al., 2014; Jordan et al., 2013; Van Hoof et al., 2017). The only study to our knowledge to collect both number line and whole number comparison found that number line estimation was a stronger predictor of rational number understanding than whole number comparison (Van Hoof et al., 2017). Number line tasks may better capture learners' understanding of numbers as quantities that represent magnitudes, whereas comparison tasks may index automatic access to numbers' ordinal information (Lyons & Beilock, 2013). This view of number comparison may explain its contribution to same numerator problems, where instead of interfering with these supposedly contradictory problems, it indexes automatic access to ordinal information that can be used to quickly identify the *smaller* quantity. Although performance on whole number comparison and number line estimation tasks tend to be positively correlated (Laski & Siegler,

2007), further research should include both measures in order to disentangle which aspects of whole number knowledge support rational number outcomes.

The role of inhibitory control beyond whole number interference

In the current study, we sought to test to predictions of the whole number interference hypothesis, namely that one source of difficulty with rational numbers is interference caused by whole number magnitude knowledge. One prediction from this hypothesis is that whole number knowledge should hinder fraction comparison performance because many properties of whole numbers do not apply to rational numbers. However, we did not observe a negative relation between whole number knowledge and fraction comparison performance, and furthermore, we did not find evidence for an interaction between whole number knowledge and inhibitory control. In other words, the participants with better whole number knowledge did not need more inhibitory control to overcome that knowledge. This pattern of results suggests that although inhibitory control contributes to fraction comparison performance, it is serving a function other than inhibiting whole number knowledge.

The performance patterns for the congruent versus incongruent problems with distinct components (Level 2) provides some insight into alternative contributions of inhibitory control. We hypothesized that congruency with whole number knowledge would be a salient factor that influences fraction comparison performance for problems with distinct components. By this logic, incongruent comparisons should be more difficult than congruent ones because they involve two sub-comparisons that contradict whole number knowledge (e.g., in the comparison of $\frac{2}{3}$ vs. $\frac{5}{9}$, the larger magnitude fraction contains both the smaller numerator and denominator components). Contrary to our expectations, participants were more accurate on the incongruent distinct component problems than on the congruent problems. However, a number of previous studies have documented stronger performance for incongruent relative to congruent problems with distinct components (D. M. Gómez & Dartnell, 2018; González-Forte et al., 2018, 2020; Obersteiner et al., 2013; Rinne et al., 2017; Toledo et al., 2022), which may reflect the strategy that participants engaged in when solving these problems. In these cases, participants (or at least a subset of participants), may have used a “select the smaller number strategy,” which would lead to the correct answer for incongruent but not congruent problems. In fact, recent studies have identified this exact strategy among a subset of children (D. M. Gómez & Dartnell, 2018; Miller Singley & Bunge, 2018; Rinne et al., 2017). This strategy may reflect a transition from a more naïve “select the larger strategy” to a more sophisticated understanding of fractions, whereby students note that something is “different” about fractions and acknowledge that the presence of larger numbers is not necessarily an indicator of a larger magnitude (Rinne et al., 2017). In turn, students who have this understanding may be monitoring their responses more carefully, thus invoking inhibitory control for both congruency problem types. An important direction for future studies will be to move beyond simple accuracy measures and instead use participants’ individual patterns of success and failure across problem types to identify different underlying strategies or learning profiles (Braithwaite et al., 2019).

Another possible source of performance differences between congruent and incongruent comparisons is in the properties of the fraction pairs. Constructing well-matched sets of

congruent and incongruent fraction comparison pairs is difficult because controlling for one feature sometimes makes it mathematically difficult to control for another feature (Rosenberg-Lee, 2021). In the present study, we chose to control for magnitude difference between the fraction pairs and partial distance. Although, we were not able to exactly match between the conditions on factors like gap distance, half benchmarking, simplified forms, and familiarity, which have been shown to impact performance (D. M. Clarke & Roche, 2009; González-Forte et al., 2018, 2020), there does not seem to be a definitive impact of these factors on students' performance in the current sample (Table 2.2.S3). For example, applying the gap strategy (selecting the fraction with the smallest distance between numerator and denominator) would only lead to an error on one congruent pair (2/7 vs 3/9), yet performance was no worse on this pair (see Table 2.2.S3). Future studies should systematically manipulate these factors to explicitly test how inhibitory control relates to these stimulus features. It is also important to note that the fraction task used in this study was timed, whereas children and adults do not often have stringent time constraints in most real-world situations with fractions. Even though participants were not using the full RT window available, students may use different strategies when under time pressure compared to when they can pace themselves. Thus, another future direction is to compare the role of inhibitory control in timed and untimed fraction tasks.

Beyond the specifics of the fraction stimuli, the pattern of inhibitory control engagement found here is in line with prior work that suggests that learning about rational numbers requires conceptual change (McMullen et al., 2015; McNeil & Alibali, 2005; Vamvakoussi & Vosniadou, 2004), and that conceptual change invokes inhibitory control (Bascandziev et al., 2018; Brookman-Byrne et al., 2018). Specifically, in cases in which learners acquire new counter-intuitive knowledge that contradicts previous knowledge, learners maintain both conceptual frameworks and inhibitory control is required to activate the correct knowledge and inhibit the initial, inappropriate knowledge (Shtulman & Valcarcel, 2012; Vosniadou, 2014). In the case of fractions, what therefore needs to be inhibited may not be specifically knowledge of whole number magnitudes, but rather the entire conceptual framework of number properties that apply to whole numbers but not to rational numbers.

Conclusions

Consistent with a more general proposal that overcoming interference is a key building block for mastering rational numbers (Rosenberg-Lee, 2021), we found that individual differences in inhibitory control predicted children's fraction comparison performance. Further, this effect was independent of the contribution of working memory. However, contrary to the counterintuitive prediction about the role of whole number knowledge, we did not find that superior whole number knowledge hindered fraction understanding. Instead, whole number knowledge positively predicted performance for the easier, shared component problems and was not related to performance on distinct component problems. Further, we found no differences in these patterns between congruent and incongruent problems, and developmentally, the contributions of these factors were generally stable from 3rd to 7th grade. Together, these results converge with the growing body of evidence pointing to a role for inhibitory control in rational number understanding (Abreu-Mendoza et al., 2020; Avgerinou & Tolmie, 2019; Coulanges et al., 2021; D. M. Gómez et al., 2015; Ren & Gunderson, 2021), and further suggest that its contribution may be

more general than overcoming whole number magnitude knowledge. Given that individual differences in inhibitory control are evident far before children are exposed to formal fraction instruction, assessing this capacity early on could be useful to identify students who are likely to benefit from additional support while learning fractions.

2.3 Relational thinking: An overlooked component of executive functioning

Abstract

Relational thinking, the ability to represent abstract, generalizable relations, is a core component of reasoning and human cognition. Relational thinking contributes to fluid reasoning and academic achievement, particularly in the domain of math. However, due to the complex nature of many fluid reasoning tasks, it has been difficult to determine the degree to which relational thinking has a separable role from cognitive processes collectively known as executive functions (EFs). Here, we used a simplified reasoning task to better understand how relational thinking contributes to math achievement in a large, diverse sample of elementary and middle school students (N = 942). Students also performed a set of ten adaptive EF assessments, as well as tests of math fluency and fraction magnitude comparison. We found that relational thinking was significantly correlated with each of the three EF composite scores previously derived from this dataset, albeit no more strongly than they were with each other. Further, relational thinking predicted unique variance in students' math fluency and fraction magnitude comparison scores over and above the three EF composites. Thus, we propose that relational thinking be considered an EF in its own right as one of the core mid-level cognitive abilities that supports cognition and goal-directed behavior.

Introduction

Relational thinking, or the process of identifying and integrating relations, is regularly invoked during reasoning (Doumas et al., 2008). Among other things, it enables us to draw higher-order abstractions and generalize across situations and contexts (Gentner, 2003). Relational thinking is central to measures of fluid reasoning, and the terms reasoning and non-verbal intelligence are sometimes used interchangeably to describe aspects of intelligence that are separable from crystallized intelligence (P. A. Carpenter et al., 1990; Cattell, 1987a). Although some other animals can represent abstract relations between items, such as *same* and *different*, humans are unparalleled with respect to the ability to consider and integrate relations (Gentner et al., 2021; Penn et al., 2008; R. K. R. Thompson & Oden, 2000). For example, we can use analogical reasoning to intuit that the relation between a hand and a glove and is the same as that between a foot and a sock. Likewise, we can use transitive inference to deduce that if a cat is bigger than a squirrel and a squirrel is bigger than a mouse, then a cat is bigger than a mouse. Here, we argue that relational thinking is a core cognitive ability that should be considered an executive function (EF).

EFs are construed as a constellation of domain-general, effortful cognitive processes that are critical for goal-directed behavior (Diamond, 2013), flexible thinking and problem solving (Cragg & Gilmore, 2014; Lehto et al., 2003), reasoning (Richland et al., 2006; Richland & Burchinal, 2013), and, as a result, academic performance (Best et al., 2011; Lawson & Farah, 2015; Rose et al., 2011). As such, they can be thought of as mid-level cognitive processes situated between basic perceptual, attentional, and motor processes, on the one hand, and high-level cognitive abilities (e.g., language, reading, math) on the other. Developmental psychology research on EFs commonly focuses on three putative core abilities: inhibition (the ability to selectively control attention and resist interference), working memory (the ability to hold, update, and manipulate information in mind), and cognitive flexibility (the ability to switch between perspectives, rules,

and schemas as needed on a moment-to-moment basis; also referred to as shifting) (Diamond, 2013; Lehto et al., 2003; Miyake et al., 2000; Rose et al., 2011). These three EFs are theorized to be distinct abilities that frequently interact to support high-level cognition and behavior. Many previous studies on the development and structure of EFs have focused on this hypothesized structure and chosen tasks that map onto these components (Hughes et al., 2009; Huizinga et al., 2006; K. Lee et al., 2013; Lehto et al., 2003).

Even within these three canonical components of EF, however, there is variability in how each is conceptualized. For example, the inhibition construct combines inhibition of attention and inhibition of action, though these two types of inhibition may be cognitively and neurally distinct (Bunge et al., 2002; Diamond, 2013). Furthermore, there is no single gold-standard definition of what makes a cognitive ability an EF, and the EF components found in any particular study are directly related to the selected tasks. More broadly, there is no agreed upon taxonomy of EFs across psychology and neuroscience.

Like the canonical EFs, relational thinking has long been viewed as a domain-general, effortful, mid-level cognitive process that is central to higher-level human cognition—in particular, various forms of reasoning (Alexander, 2016; Cattell, 1987a; Halford et al., 2010). Despite this parallel conceptualization, relational thinking tasks have not been included in studies assessing the structure of EFs. As a result, these standard models do not involve a relational component. However, this historical precedent in and of itself does not mean that relational thinking should not be considered an EF.

To tackle the question of whether relational thinking should be conceptualized as an EF, it is important to test whether it is distinct from the canonical EFs. Many previous studies have noted a relation between reasoning abilities and EFs (e.g., Conway et al., 2003; Duncan et al., 2012; Engle et al., 1999; Friedman et al., 2006; Van Aken et al., 2016; see Diamond, 2013 for a review). Broadly, these studies demonstrate that individuals who score higher on standard measures of EFs also tend to score higher on standard measures of reasoning. In particular, working memory and inhibitory control are frequently found to correlate positively with reasoning in both adults (Grossnickle et al., 2016; Krawczyk et al., 2008) and children (Fry & Hale, 2000; Richland et al., 2006; Richland & Burchinal, 2013; Starr et al., 2018; Thibaut et al., 2010; Thibaut & French, 2016).

The tight relation between reasoning and EFs has led some researchers to conclude that reasoning is not actually a separable ability from EFs (Martínez et al., 2011), whereas others have suggested that EFs explain only about half of the variance in reasoning ability (Friedman et al., 2006). These arguments have been clouded by the fact that most measures of reasoning are complex and engage canonical EFs as well as relational thinking, which makes it difficult to determine whether this overlap stems from confounds in tasks themselves versus a true association between the underlying abilities.

Breaking down the steps required to solve classic matrix reasoning tasks (Cattell, 1940; Raven, 1941, 2000) highlights how both relational thinking and canonical EF abilities are required for success. Matrix reasoning tasks are frequently used as a stand-alone fluid reasoning task but are also a typical component of nonverbal intelligence tests (e.g., the Wechsler Adult Intelligence

Scale; Wechsler, 1958). In this type of task, participants are shown a grid of visuospatial designs with one item missing, and participants must choose the correct item to complete the pattern from an array of choice items. For example, the grid may be organized such that rows of items vary along one relation (e.g., size) and the columns vary along another relation (e.g., shape). Selecting the correct item that completes the grid requires integrating the two separate relations in order to identify the item that completes the pattern in both dimensions. Therefore, in addition to identifying and integrating relations, the task also engages working memory to maintain and manipulate the different critical relations in the focus of attention and inhibitory control to resist selecting salient distractor items (Chen et al., 2016; Sternberg, 1977; Stevenson & Hickendorff, 2018; Vodegel Matzen et al., 1994). Other complex reasoning tasks similarly tap both relational thinking and canonical EFs (Richland et al., 2006; Richland & Morrison, 2010; Starr et al., 2018; Thibaut et al., 2010). Therefore, an important first step for understanding how relational thinking relates to canonical EFs is to design a task that engages relational thinking while minimizing demands on other EFs.

EFs are often described as processes that support academic achievement; thus, exploring the unique contribution of relational thinking to math performance is germane to the question of whether it should be considered an EF. EFs are strong predictors of academic achievement throughout childhood and adolescence (Best et al., 2011; Cowan, 2014; Ferrer et al., 2007; Ferrer & McArdle, 2004; Richland et al., 2007; St Clair-Thompson & Gathercole, 2006). This relation has been particularly well-documented in the domain of mathematics: children who score higher on EF assessments also typically score higher on lab-based and school-based math assessments (Bull & Scerif, 2001; Fuchs et al., 2012; Geary, 2011; Green et al., 2017; Purpura & Ganley, 2014; Richland et al., 2007; St Clair-Thompson & Gathercole, 2006; Taub et al., 2008). Some of the ways EFs support math achievement include helping learners maintain relevant knowledge in mind (working memory), inhibit inappropriate strategies (inhibitory control), and switching between different strategies (cognitive flexibility) (Bull & Lee, 2014; Cragg & Gilmore, 2014).

In parallel to the research linking canonical EFs to math achievement, a number of studies have demonstrated that reasoning ability predicts both current and future math abilities (Fuchs et al., 2006; Green et al., 2017; Taub et al., 2008). Mathematical thinking is inherently relational (DeWolf et al., 2015; Miller Singley & Bunge, 2018; Richland et al., 2007). Students who conceptualize math as a relational system are more successful in tackling novel problem types and formats than students who conceptualize math as a set of explicit rules and procedures (Richland et al., 2012). The importance of relational thinking for math is particularly evident for concepts like equivalence, algebra, and fractions. With respect to fractions, for example, the magnitude of a fraction is equivalent to the relation between the numerator and the denominator, and comparing fractions therefore requires integrating the relation between one numerator and denominator with the relation between the other numerator and denominator (Bonato et al., 2007; Miller Singley & Bunge, 2014). Previous studies have found associations between children's performance on fraction comparison tasks and both reasoning and EF measures (DeWolf et al., 2015; Hecht et al., 2003; Kalra et al., 2020; Miller Singley & Bunge, 2018; Siegler & Pyke, 2013).

However, as previously described, the complex nature of many reasoning tasks muddles the interpretation of these relations between EFs, reasoning, and math achievement. Based on the current research, it is unclear whether relational thinking and canonical EFs are explaining unique or overlapping variance in children's math achievement. Clarifying the relation between relational thinking, EFs, and math is necessary to better understand the foundational skills required for students to succeed with mathematical concepts.

The goals of the present study were threefold. First, we aimed to assess age-related differences and individual variability on a simplified relational thinking task in a large developmental sample. Second, we sought to examine the degree to which relational thinking can be considered a separable construct from EFs that are thought to contribute to reasoning, namely working memory and inhibitory control. Finally, we further explored the hypothesis that relational thinking should be considered a distinct EF by testing whether it independently contributes to math performance, over and above common measures of EF.

The relational thinking task used here is a form of relational match-to-sample task (Christie & Gentner, 2014; Christoff et al., 2003; Premack, 1983; L. B. Smith, 1984; R. K. R. Thompson & Oden, 2000). It takes the form of proportional analogies (A:B::C:D), but has no semantic component and therefore requires little or no prior knowledge (Figure 2.3.1). Briefly, this task requires participants to jointly consider the relations between two pairs of simple visual stimuli that vary along either two or three dimensions in order to determine whether the two pairs share the same relation (e.g., both match in shape). We created two levels of relational complexity (Halford et al., 1998) within this task. The first level requires consideration of two features, as in prior work (Christoff et al., 2003; Dumontheil et al., 2010; Wendelken et al., 2011); the second is a novel, more challenging level that requires consideration of three features. We designed this measure such that it can be administered efficiently in a group setting for use with large-scale data collection (Uncapher, 2018).

This task explicitly taps relational thinking skills, and there are critical differences in its design that reduce demand on EFs in comparison to standard reasoning tasks. First, the task contains only four elements and a two-alternative forced-choice answer structure to reduce demands on working memory and inhibitory control. On each trial, participants must decide only whether the items do or do not match, rather than choosing between up to eight answer choices. Second, participants are told the rules in advance and must only follow those two or three rules (depending on the level of relational complexity), rather than needing to induce multiple novel rules on their own (P. A. Carpenter et al., 1990). Finally, the task involves only a small set of geometric shapes, rather than familiar real-world objects, to minimize the involvement of semantic knowledge. By reducing the number of elements—both sample and choice items—involved in the task, specifying a limited set of rules in advance, and using basic shapes, this relational thinking task is designed to isolate relational thinking while minimizing the involvement of other EFs. However, we do not consider it possible to fully eliminate demands on canonical EFs in any task that requires rule-guided behavior (Bunge & Zelazo, 2006)—any more than we consider it possible to devise a "pure" EF task, as even the canonical EFs are theorized to interact with one another (Blackwell et al., 2014; Diamond, 2013).

We addressed our aims in the context of a large, longitudinal investigation of the development of EF components across Grades 3-8. Participants in this study completed a battery of nine EF tasks from the Adaptive Cognitive Evaluation (ACE) battery (Younger et al., 2022). A primary analysis of the ACE data, which used factor analysis methods to uncover how the different tasks grouped together, were published separately (Younger et al., 2023). Based on these analyses, we used composite scores to index three putative EFs: working memory, interference resolution and response inhibition (different forms of inhibitory control, involving suppression of visual distractors and motoric responses, respectively), as well as a single measure of cognitive flexibility. In addition, participants completed the relational thinking task and a battery of scholastic achievement tasks that included tests of math fluency and fraction comparison. The present series of analyses relate children's relational thinking task performance to individual differences in the three EF composite scores, cognitive flexibility, math fluency, and fraction comparison scores. We focus on children in 4th, 6th, and 8th grades in order to investigate how the relations between these different abilities change between elementary and middle school.

Method

Participants

Participants in the present study were part of Project iLead, a two-year, multi-site study investigating EF development throughout elementary and middle school (Younger et al., 2022). Data collection took place at nine schools in northern California. In total, 1,280 students participated over the course of two years. The data described here come from year two of the study, which included 288 fourth graders, 336 sixth graders, and 482 eighth graders. Of those participants, 243 fourth graders, 270 sixth graders, and 429 eighth graders had valid data for the relational thinking task and were included in our analyses. The demographic characteristics of our sample are detailed in Table 2.3.1.

The study was performed in accordance with protocols approved by the Institutional Review Board (IRB) of the University of California, San Francisco. Written parental or guardian consent was obtained from all participants at the beginning of the study, and verbal assent from all participants was obtained before all in-class data collection sessions. At the end of the study, all students in participating classrooms received snacks and stickers, regardless of their individual participation.

Procedure

Participants were tested during school hours at the beginning and end of each academic year (fall and spring) over two academic school years. EF and math fluency assessments occurred at all four time points. The reasoning and fraction tasks were part of one of two scholastic assessments that were administered to participants once per year in alternating semesters. Students were randomly assigned to complete each task set in either the fall or spring of each year. At each of the four timepoints, the EF assessments were administered first, and the scholastic assessments were administered approximately six weeks later ($M = 5.7$ weeks, $SD = 2.4$, min. = 1.9, max. = 10).

All tasks were administered in a group setting on iPads. Each group administration was conducted by 4-12 researchers, in proportion to the student group size. A lead facilitator gave verbal

Table 2.3.1. Demographic characteristics of sample. Because some parents opted not to share demographic data for their children, not all columns sum to 100%.

Variable	Grade 4	Grade 6	Grade 8
Age (years)	9.81 [8.94, 11.90]	11.75 [10.60, 13.26]	13.76 [12.87, 15.35]
Gender			
Female	53%	49%	50%
Male	47%	51%	50%
Ethnicity			
American Indian or Alaskan Native	0%	0%	1.2%
Asian	50%	40%	35%
Black or African American	0.9%	2.5%	1.2%
Blank on Purpose	0.5%	0%	0.2%
Filipino	3.3%	6.2%	7.9%
Hispanic or Latino	19%	28%	32%
Pacific Islander	0.5%	0.4%	0.2%
Two or More Races	7.0%	3.7%	4.7%
White or Caucasian	19%	19%	18%
Eligible for Free or Reduced Lunch	28%	35%	35%

instructions to the group for each task, aided by visual instructions from a 24" x 36" flipbook. Participants began each task at the same time, and instructions for the next task were not given until all participants completed the current task. Each task began with practice trials during which researchers monitored participants to ensure participants understood the task and were correctly following task instructions. Researchers monitored the sessions throughout administration to provide technical assistance, answer student questions, and monitor performance. Administration sessions lasted approximately 50 minutes.

Relational thinking task

The relational thinking task (Figure 2.3.1) was adapted from a task that has been used previously to study the neural correlates of relational reasoning and its development (Christoff et al., 2003; Dumontheil et al., 2010; Wendelken et al., 2011). The experimenter introduced the game by telling participants that in this game, “We want to see if the top row matches the bottom row using the same rule.” On each trial, participants saw two pairs of items. In Level 1, the items varied in both color and shape. Participants decided whether the pairs matched along the same dimension (i.e., in both pairs, the items within each pair both matched or shape or color). If the items did match along the same direction, the participant was to press a button marked “YES” on

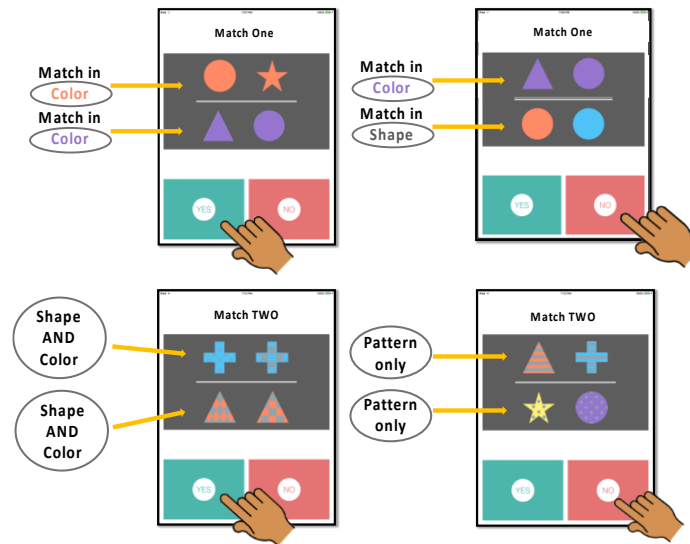


Figure 2.3.1. Four sample trials from the relational thinking task. In Level 1 (top), both pairs of images must match in the same dimension to be classified as a match. In Level 2 (bottom), both pairs of images must match in two dimensions to be classified as a match. Participants must identify the dimensions along which items match; labels are shown in the figure for illustrative purposes only. The hands, also for illustrative purposes only, indicate the correct response option for each trial.

the screen. If the items did not match along the same dimension, the participant was to press a button marked “NO” on the screen. Level 2 added the dimension of pattern: participants needed to decide whether the top and bottom pairs both followed the same two matching rules (i.e., they matched in shape and color, shape and pattern, or color and pattern). For this level, explicit instructions and practice trials made it clear that the pairs needed to match in two dimensions, and that pairs that matched in only one dimension were not matches. Participants again responded by pressing the “YES” or “NO” buttons on the screen.

In each level, half of the trials represented matches and half of the trials were non-matches. The order of the trials was randomized. The response window was 3.5 seconds in Level 1, and 4.5 seconds in Level 2. Participants completed three practice trials of Level 1 as a group and four practice trials individually with feedback before completing 20 Level 1 test trials without feedback. Next participants completed five Level 2 practice trials as a group and four practice trials individually, with feedback. However, only participants who achieved at least 75% accuracy on Level 1 moved onto the Level 2 test trials. Participants who scored below 75% completed Level 1 again, but only scores from the first round of gameplay were analyzed.

Math fluency task

Math fluency was measured using an assessment that tested participants’ ability to quickly and accurately answer math problems, similar to the Math Fluency task of the Woodcock-Johnson III Tests of Achievement (Schrank et al., 2014). The assessment required participants to solve single-digit math equations (addition, subtraction, and multiplication) by typing the correct answer. Math equations were presented one at a time, and the task would not advance until a response was made for each trial. Participants were asked to solve as many equations as they were able in

three minutes. Two practice trials were administered to ensure understanding. Scores were determined by the total number of correct responses.

Fraction comparison task

Proficiency with fractions was measured using an assessment that tested participants' ability to compare numerical magnitudes quickly and accurately. The task had three levels. Level 1 required participants to compare symbolic digits, and Levels 2 and 3 required them to compare symbolic fractions. In Levels 2 and 3, each trial presented two single-digit fractions, side-by-side, and participants were instructed to indicate which fraction magnitude was larger. In Level 2, both fractions within a trial shared the same numerator or denominator (e.g., $2/3$ vs. $1/3$). In Level 3, the numerators and denominators always differed (e.g., $5/6$ vs. $3/7$). Participants responded by touching the larger fraction. Participants completed 16 trials each in Levels 2 and 3, with a response window of 4.5 s. Participants needed to answer at least 75% of the trials accurately in a given level to advance to the next level. Only data from Level 2 were included in the present analyses, because Level 1 did not involve fractions and few 4th grade students progressed to Level 3.

Executive function tasks

EFs were assessed with the Adaptive Cognitive Evaluation (ACE), an iPad-based battery that assesses EF skills and is composed of ten tasks developed from commonly used EF assessments: basic response time, forward spatial span, backward spatial span, impulsive attention, sustained attention, tap and trace, color-word Stroop, letter flanker, boxed, and task switch (Table 2.3.2; Younger et al., 2022). A full description of the ACE tasks and its adaptive algorithm can be found in Younger et al. (2022).

Data from all four timepoints were previously analyzed using explanatory and confirmatory analysis methods to determine the underlying organization (Younger et al., 2023). These analyses revealed that a three-factor model of EFs fit the data for all cohorts at all four timepoints. The three components were labeled response inhibition (sustained attention, impulsive attention, and tap and trace), interference resolution (Stroop, flanker, boxed), and working memory (forward and backward spatial span). Because the factor loadings varied slightly between cohorts and timepoints, we calculated composite scores for each factor by z-scoring the individual task scores for each cohort and timepoint and then averaging the z-scores for the tasks that comprised each factor. We used the three composite scores to index three EFs². In addition, as a measure of cognitive flexibility we used stand-alone, z-scored scores from the task switching task because a technical error at the first timepoint prevented scores from this task from being used in the factor analyses.

² Because the composite EF scores do not preserve subtle difference in factor loadings across the cohorts and time points, we also ran all of the analyses with factor scores instead of composite scores. The main pattern of results remained unchanged. Because the analyses using the true factor scores and the simplified composite scores lead to the same conclusions about the role of relational thinking and its association to the other EFs, we report the composite scores for coherence with other manuscripts based on this dataset.

Table 2.3.2. Overview of EF tasks in the ACE battery and labels provided for the EF composites derived from these tasks in the parent study (Younger et al., 2022, 2023).

Task Name	Description	Theorized Construct	EF Composite
Basic Reaction Time	Tap in response to visual targets	Processing speed	N/A (regressed from performance metrics of all other tasks to control for general differences in processing speed)
Forward Spatial Span	Tap to recreate cued spatial sequence of targets	Working memory	Working memory
Backward Spatial Span	Tap to recreate cued spatial sequence of targets in reverse order	Working memory	Working memory
Impulsive Attention	Respond to frequent targets and withhold response to non-targets	Inhibitory control	Response inhibition
Sustained Attention	Respond to infrequent targets and withhold response to frequent non-targets	Sustained attention	Response inhibition
Tap and Trace	Tap with dominant hand and trace shapes with the non-dominant hand	Dual-task performance	Response inhibition
Stroop	Respond to text colors that are congruent or incongruent with semantic meaning	Inhibitory control	Interference resolution
Flanker	Respond to middle letters that are congruent or incongruent with flanking letters	Selective attention; Inhibitory control	Interference resolution
Boxed	Identify target stimuli within arrays of distractor stimuli	Visual search	Interference resolution
Task Switch	Switch between responding to color or shape of target stimuli in response to pre-trial cues	Cognitive flexibility	N/A (technical error prevented inclusion in factor analyses)

Data analysis

Data from the relational thinking and fraction comparison tasks were first cleaned by removing trials with response times that fell more than three median absolute deviations (MAD) above or below each participant’s median response time (Leys et al., 2013). This removed approximately 3% of trials from relational thinking Levels 1 and 2 and approximately 4% of fraction comparison trials. For each task and level, participants needed to have valid response data (i.e., a response was recorded within 3 MAD of their median response time) on at least 2/3 of trials in order to be

included in further analysis. Eleven participants in the relational thinking Level 1 task, 3 participants in the relational thinking Level 2 task, and 5 participants in the fraction comparison task did not have enough valid trials and were excluded. All data were analyzed in R. We used the lmerTest package (Kuznetsova et al., 2017) to conduct mixed-effects models and compared successive models using the anova function from the base stats package (R Core Team, 2013). We used the Raincloud package (Allen et al., 2021) to visualize group performance.

Results

Relational thinking task performance

The first series of analyses examined performance on each level of the relational thinking task. Overall, all grade levels performed above chance on Level 1 (see Table 2.3.3 for a summary of performance on the relational thinking and math tasks). A logistic mixed-effects model predicting trial accuracy with grade, semester in which testing occurred, and number of previous testing sessions as fixed effects and school and participant as random effects revealed that accuracy increased with grade level ($\beta = 0.15$, $SE = 0.07$, $p < 0.001$; Figure 2.3.2A). A similar linear model predicting response times (RTs) also showed RTs became faster as function of grade level ($\beta = -32.77$, $SE = 26.16$, $p < 0.001$; Figure 2.3.2A). These results suggest that relational thinking continues to improve throughout elementary and middle school.

Next, we examined the data from Level 2. Only participants who achieved at least 75% accuracy on Level 1 moved onto Level 2. By this criterion, 75/243 4th graders (31%), 145/270 6th graders (54%), and 266/431 8th graders (62%) moved onto Level 2. A chi-squared test for trend in proportions indicated that the proportion of participants advancing to Level 2 increased with grade ($\chi^2 = 55.82$, $p < .001$). Similar grade-wise developmental trends were found for Level 2 performance. A logistic mixed-effects model predicting trial accuracy with grade, semester in which testing occurred, and number of gameplays as fixed effects and school, and participant as random effects revealed that accuracy increased with grade level ($\beta = 0.07$, $SE = 0.02$, $p < 0.001$; Figure 2.3.2B). A similar linear model predicting RTs, however, found that participants did not become significantly faster with grade ($\beta = -47.61$, $SE = 58.8$, $p = 0.42$; Figure 2.3.2B). These analyses demonstrate that, overall, both the simpler and more complex forms of relational thinking assessed by our task continue to show developmental improvements through the end of middle school.

Relations between relational thinking and EFs

In the next series of analyses, we investigated how the EF measures relate to accuracy on Level 1 of the relational thinking task in each grade. We used data from only Level 1 in these—and all subsequent—analyses because it enabled us to include a larger proportion of our sample, particularly for the 4th grade students. Another good reason for limiting subsequent analyses to Level 1 is that we assume it places fewer demands on other EFs than Level 2.

As shown in Figure 2.3.3, accuracy on the relational thinking task is significantly correlated with all three of the EF composites and task switching in 4th grade, correlated with working memory and task switching in 6th grade, and correlated with working memory, interference resolution, and task switching in 8th grade. Notably, the correlations between relational thinking and each of the

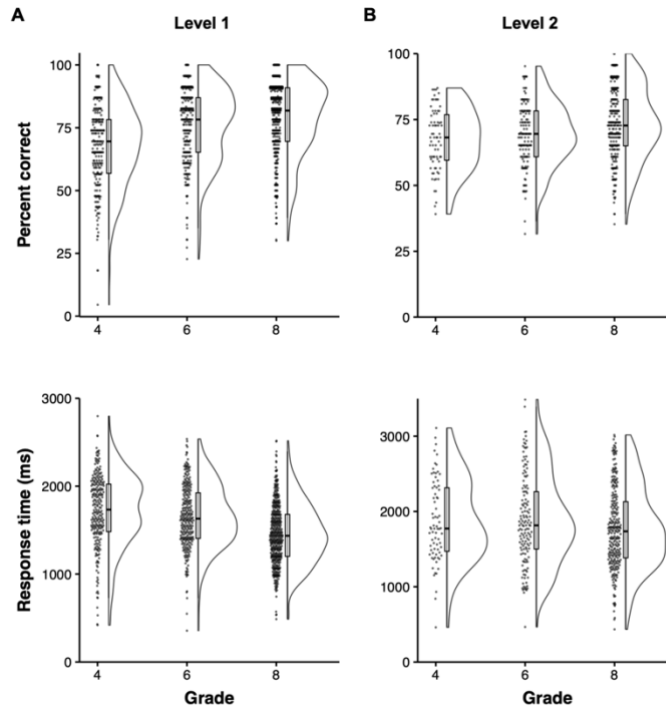


Figure 2.3.2. Accuracy and response time distributions by grade for (A) Level 1 and (B) Level 2 of the relational thinking task.

Table 2.3.3. Average performance by grade on the relational thinking task Levels 1 and 2, math fluency, and fraction comparison tasks (mean score and range for each assessment).

Measure	Grade 4	Grade 6	Grade 8
Relational Thinking Level 1 Accuracy (% correct)	68 [5, 100]	76 [23, 100]	79 [30, 100]
Relational Thinking Level 1 RT (ms) on correct trials	1,719 [420, 2,796]	1,651 [356, 2,536]	1,453 [488, 2,515]
Relational Thinking Level 2 Accuracy (% correct)	67 [39, 87]	69 [32, 95]	73 [35, 100]
Relational Thinking Level 2 RT (ms) on correct trials	1,869 [461, 3,109]	1,880 [466, 3,488]	1,770 [434, 3,016]
Math Fluency Raw score	45 [17, 77]	53 [31, 76]	56 [0, 88]
Fraction Comparison Accuracy (% correct)	71 [18, 100]	83 [21, 100]	83 [27, 100]

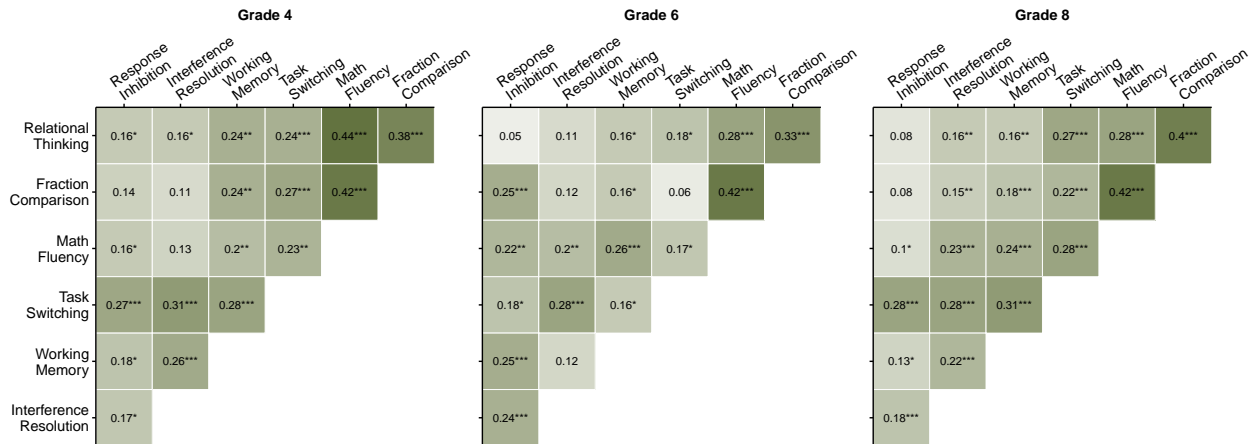


Figure 2.3.3. Pearson correlation coefficients for pairwise comparisons between all variables of interest for each grade. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; p -values are not corrected for multiple comparisons.

canonical EF measures were relatively weak at all grades (r -values ranging from .05 to .27; median r -value: .16); in fact, they were generally lower than the correlations among the canonical EF measures themselves (r -values from .17 to .31; median: .25).

Next, for each grade, we tested whether relational thinking can be predicted based on the EF composite and task switch scores. We compared linear mixed effects models that predicted relational thinking based on the semester in which participants were tested and the number of times they had completed the relational thinking task, as well as a random effect of school (Model 1), to a model that additionally included the EF composite and task switch scores (Model 2). If the model including EFs was the superior model, as indicated by an ANOVA test, we then examined the coefficients of each EF score.

In each cohort, the model containing the EF scores (Model 2) was superior to the model without these scores (4th grade: $\chi^2 = 12.97$, $p < .001$; 6th grade: $\chi^2 = 12.06$, $p < .001$; 8th grade: $\chi^2 = 29.87$, $p < .001$; Tables 2.3.4–6). However, different EF measures were predictive of relational thinking at each grade level. In 4th grade, no EF measure individually predicted unique variance. In 6th grade, working memory was the only significant predictor; in 8th grade, task switching was the only significant predictor. Thus, although relational thinking is correlated with other EF measures in this dataset, the relation between these different metrics of cognitive functioning is not stable over time and is relatively weak—certainly no higher than the relations among the canonical EF measures.

Relations between cognitive variables and math performance

Math fluency. Next, we investigated the relative contributions of relational thinking and canonical EFs to students' math fluency scores. In each cohort, we first predicted math fluency scores from a base model consisting of testing semester, the number of times the participant had previously seen the task, and a random effect of school (Model 1). Then we compared the base model to one that additionally included the EF scores as predictors (Model 2) and one that included both the EF scores and relational thinking (Model 3). The model with relational thinking

Table 2.3.4. Model coefficients for models predicting 4th grade relational thinking scores from EF scores

Predictor	Model 1	Model 2
2 Previous Sessions	0.56 ** [0.21, 0.91]	0.38 * [0.02, 0.73]
3 Previous Sessions	-0.01 [-0.86, 0.84]	-0.23 [-1.07, 0.61]
Semester	-0.15 [-0.49, 0.18]	-0.13 [-0.45, 0.19]
Response Inhibition		0.12 [-0.12, 0.37]
Interference Resolution		0.06 [-0.17, 0.29]
Working Memory		0.17 [-0.01, 0.36]
Task Switching		0.13 [-0.03, 0.29]
N	188	188
N (school)	7	7
AIC	536.72	543.36
BIC	556.14	575.73
R ²	0.10	0.14

Note. All continuous predictors are mean-centered and scaled by 1 standard deviation. Beta coefficients are standardized, and bracketed values indicate 95% confidence intervals. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Previous sessions: the number of times the tests had previously been administered to a participant; Semester: fall or spring; N: number of participants; N (school): number of schools.

Table 2.3.5. Model coefficients for models predicting 6th grade relational thinking scores from EF scores

Predictor	Model 1	Model 2
2 Previous Sessions	0.09 [-0.22, 0.39]	0.10 [-0.22, 0.41]
Semester	-0.13 [-0.40, 0.15]	-0.05 [-0.33, 0.22]
Response Inhibition		-0.03 [-0.28, 0.21]
Interference Resolution		0.15 [-0.15, 0.46]
Working Memory		0.18 * [0.01, 0.36]
Task Switching		0.14 [0.00, 0.28]
N	213	213
N (school)	3	3
AIC	619.92	626.26
BIC	636.73	656.52
R ²	0.02	0.09

Note. All continuous predictors are mean-centered and scaled by 1 standard deviation. Beta coefficients are standardized, and bracketed values indicate 95% confidence intervals. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Previous sessions: the number of times the tests had previously been administered to a participant. Semester: fall or spring. N: number of participants. N (school): number of schools.

Table 2.3.6. Model coefficients for models predicting 8th grade relational thinking scores from EF scores

Predictor	Model 1	Model 2
2 Previous Sessions	0.14 [-0.22, 0.50]	0.23 [-0.13, 0.58]
3 Previous Sessions	0.90 [-0.26, 2.06]	1.04 [-0.09, 2.16]
Semester	0.32 ** [0.11, 0.52]	0.22 * [0.02, 0.42]
Response Inhibition		-0.06 [-0.23, 0.10]
Interference Resolution		0.12 [-0.07, 0.32]
Working Memory		0.11 [-0.02, 0.23]
Task Switching		0.22 *** [0.11, 0.33]
N	368	368
N (school)	3	3
AIC	1043.54	1035.63
BIC	1066.99	1074.71
R ²	0.08	0.14

Note. All continuous predictors are mean-centered and scaled by 1 standard deviation. Beta coefficients are standardized, and bracketed values indicate 95% confidence intervals. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Previous sessions: the number of times the tests had previously been administered to a participant. Semester: fall or spring. N: number of participants. N (school): number of schools.

predicted the most variance in each cohort (4th grade: $\chi^2 = 27.6$, $p < .001$; 6th grade: $\chi^2 = 12.26$, $p < .001$; 8th grade: $\chi^2 = 9.75$, $p = 0.002$; Tables 2.3.7–9), and relational thinking was the only predictor that was significant across all three grades in the full model. Therefore, we found that performance on the relational thinking test is a unique predictor of student’s math fluency scores after controlling for the variance explained by four metrics of canonical EFs derived from ten EF tasks.

Fraction comparison. In the final series of analyses, we investigated the role of relational thinking on students’ fraction comparison task performance. In particular, we asked whether relational thinking would predict additional variance in fraction performance after accounting for math fluency and canonical EFs. Note that in these models we did not include the random effect of school because the models failed to converge. Including school as a fixed effect did not improve the model fits, so this variable was excluded altogether. We began with a base model that predicted fraction performance from math fluency scores, testing semester, and the number of previous testing sessions (Model 1). We then compared the base model to one that additionally included the EF scores (Model 2), and then one that included both EF scores and relational thinking (Model 3). For all grades, the model with relational thinking predicted the most variance in fraction scores (4th grade: $F = 7.446$, $p = .007$; 6th grade: $F = 13.4$, $p < .001$; 8th grade: $F = 34.89$, $p < .001$; Tables 2.3.10–12), and relational thinking was the only significant domain-general cognitive predictor in the full model for all three grades. These results demonstrate that relational thinking predicts additional unique variance in student’s fraction scores, above and beyond the contributions of EFs and math fluency.

Table 2.3.7. Model coefficients for models predicting 4th grade math fluency scores from EF scores and relational thinking

Predictor	Model 1	Model 2	Model 3
2 Previous Sessions	0.29 [-0.05, 0.64]	0.09 [-0.26, 0.44]	-0.04 [-0.37, 0.29]
3 Previous Sessions	0.57 [-0.26, 1.40]	0.33 [-0.49, 1.14]	0.44 [-0.32, 1.21]
Semester	0.22 [-0.16, 0.60]	0.27 [-0.11, 0.64]	0.30 [-0.05, 0.64]
Response Inhibition		0.04 [-0.20, 0.28]	0.00 [-0.22, 0.22]
Interference Resolution		-0.03 [-0.25, 0.20]	-0.04 [-0.25, 0.16]
Working Memory		0.13 [-0.05, 0.30]	0.07 [-0.10, 0.23]
Task Switching		0.21 ** [0.06, 0.37]	0.16 * [0.02, 0.31]
Relational Thinking			0.36 *** [0.22, 0.49]
N	187	187	187
N (school)	7	7	7
AIC	523.63	530.15	509.40
BIC	543.02	562.46	544.94
R ²	0.22	0.30	0.36

Note. All continuous predictors are mean-centered and scaled by 1 standard deviation. Beta coefficients are standardized, and bracketed values indicate 95% confidence intervals. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Previous sessions: the number of times the tests had previously been administered to a participant. Semester: fall or spring. N: number of participants. N (school): number of schools.

Table 2.3.8. Model coefficients for models predicting 6th grade math fluency scores from EF scores and relational thinking

Predictor	Model 1	Model 2	Model 3
2 Previous Sessions	-0.16 [-0.46, 0.14]	-0.09 [-0.39, 0.21]	-0.13 [-0.42, 0.16]
Semester	0.22 [-0.06, 0.49]	0.31 * [0.06, 0.57]	0.32 * [0.07, 0.57]
Response Inhibition		0.19 [-0.04, 0.42]	0.19 [-0.03, 0.42]
Interference Resolution		0.37 * [0.09, 0.65]	0.33 * [0.05, 0.60]
Working Memory		0.29 *** [0.12, 0.45]	0.25 ** [0.08, 0.41]
Task Switching		0.10 [-0.03, 0.23]	0.08 [-0.05, 0.20]
Relational Thinking			0.22 *** [0.09, 0.34]
N	212	212	212
N (school)	3	3	3
AIC	612.50	595.61	589.90
BIC	629.28	625.82	623.47
R ²	0.03	0.22	0.24

Note. All continuous predictors are mean-centered and scaled by 1 standard deviation. Beta coefficients are standardized, and bracketed values indicate 95% confidence intervals. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Previous sessions: the number of times the tests had previously been administered to a participant. Semester: fall or spring. N: number of participants. N (school): number of schools.

Table 2.3.9. Model coefficients for models predicting 8th grade math fluency scores from EF scores and relational thinking

Predictor	Model 1	Model 2	Model 3
2 Previous Sessions	-0.02 [-0.38, 0.33]	0.05 [-0.29, 0.40]	0.02 [-0.32, 0.36]
3 Previous Sessions	-0.01 [-1.14, 1.13]	0.15 [-0.93, 1.24]	-0.00 [-1.08, 1.08]
Semester	0.54 *** [0.34, 0.73]	0.42 *** [0.23, 0.61]	0.39 *** [0.20, 0.58]
Response Inhibition		-0.03 [-0.19, 0.14]	-0.02 [-0.18, 0.14]
Interference Resolution		0.19 * [0.00, 0.38]	0.17 [-0.01, 0.36]
Working Memory		0.20 ** [0.07, 0.32]	0.18 ** [0.06, 0.30]
Task Switching		0.18 ** [0.07, 0.28]	0.14 ** [0.04, 0.25]
Relational Thinking			0.15 ** [0.06, 0.25]
N	368	368	368
N (school)	3	3	3
AIC	1028.47	1010.67	1007.49
BIC	1051.92	1049.75	1050.48
R ²	0.12	0.21	0.22

Note. All continuous predictors are mean-centered and scaled by 1 standard deviation. Beta coefficients are standardized, and bracketed values indicate 95% confidence intervals. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Previous sessions: the number of times the tests had previously been administered to a participant. Semester: fall or spring. N: number of participants. N (school): number of schools.

Table 2.3.10. Model coefficients for models predicting 4th grade fraction scores from math fluency, EF scores, and relational thinking

Predictor	Model 1	Model 2	Model 3
2 Previous Sessions	0.21 [-0.12, 0.54]	0.09 [-0.25, 0.43]	0.03 [-0.31, 0.36]
3 Previous Sessions	0.38 [-0.42, 1.17]	0.20 [-0.60, 1.00]	0.27 [-0.52, 1.05]
Semester	0.13 [-0.14, 0.40]	0.16 [-0.11, 0.43]	0.18 [-0.08, 0.45]
Math Fluency	0.41 *** [0.27, 0.54]	0.36 *** [0.22, 0.50]	0.28 *** [0.13, 0.43]
Response Inhibition		0.03 [-0.20, 0.26]	0.01 [-0.22, 0.24]
Interference Resolution		-0.04 [-0.26, 0.19]	-0.05 [-0.26, 0.17]
Working Memory		0.16 [-0.03, 0.34]	0.13 [-0.05, 0.31]
Task Switching		0.15 [-0.00, 0.30]	0.13 [-0.02, 0.28]
Relational Thinking			0.21 ** [0.06, 0.36]
N	185	185	185
R ²	0.19	0.23	0.26

Note. All continuous predictors are mean-centered and scaled by 1 standard deviation. Beta coefficients are standardized, and bracketed values indicate 95% confidence intervals. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Previous sessions: the number of times the tests had previously been administered to a participant. Semester: fall or spring. N: number of participants.

Table 2.3.11. Model coefficients for models predicting 6th grade fraction scores from math fluency, EF scores, and relational thinking

Predictor	Model 1	Model 2	Model 3
2 Previous Sessions	-0.02 [-0.28, 0.25]	0.01 [-0.25, 0.28]	-0.00 [-0.26, 0.25]
Semester	-0.12 [-0.38, 0.14]	-0.11 [-0.37, 0.15]	-0.08 [-0.33, 0.17]
Math Fluency	0.43 *** [0.31, 0.56]	0.40 *** [0.26, 0.53]	0.34 *** [0.20, 0.47]
Response Inhibition		0.28 * [0.05, 0.52]	0.30 ** [0.08, 0.53]
Interference Resolution		0.02 [-0.27, 0.30]	0.01 [-0.27, 0.28]
Working Memory		0.03 [-0.14, 0.20]	0.00 [-0.16, 0.17]
Task Switching		-0.04 [-0.18, 0.09]	-0.07 [-0.20, 0.06]
Relational Thinking			0.23 *** [0.11, 0.36]
N	211	211	211
R ²	0.18	0.21	0.26

Note. All continuous predictors are mean-centered and scaled by 1 standard deviation. Beta coefficients are standardized, and bracketed values indicate 95% confidence intervals. *** p < 0.001; ** p < 0.01; * p < 0.05. Previous sessions: the number of times the tests had previously been administered to a participant. Semester: fall or spring. N: number of participants.

Table 2.3.12. Model coefficients for models predicting 8th grade fraction scores from math fluency, EF scores, and relational thinking

Predictor	Model 1	Model 2	Model 3
2 Previous Sessions	0.47 ** [0.13, 0.81]	0.51 ** [0.17, 0.85]	0.44 ** [0.12, 0.77]
3 Previous Sessions	0.71 [-0.36, 1.78]	0.78 [-0.29, 1.86]	0.47 [-0.56, 1.50]
Semester	0.11 [-0.08, 0.30]	0.08 [-0.11, 0.27]	0.06 [-0.13, 0.24]
Math Fluency	0.41 *** [0.31, 0.50]	0.37 *** [0.26, 0.47]	0.31 *** [0.21, 0.41]
Response Inhibition		-0.02 [-0.18, 0.14]	-0.01 [-0.16, 0.15]
Interference Resolution		0.06 [-0.13, 0.25]	0.04 [-0.14, 0.21]
Working Memory		0.07 [-0.06, 0.19]	0.05 [-0.07, 0.17]
Task Switch		0.10 [-0.00, 0.21]	0.05 [-0.06, 0.15]
Relational Thinking			0.28 *** [0.19, 0.38]
N	366	366	366
R ²	0.20	0.21	0.28

Note. All continuous predictors are mean-centered and scaled by 1 standard deviation. Beta coefficients are standardized, and bracketed values indicate 95% confidence intervals. *** p < 0.001; ** p < 0.01; * p < 0.05. Previous sessions: the number of times the tests had previously been administered to a participant. Semester: fall or spring. N: number of participants.

Discussion

The primary aims of the present study were to delineate the development of relational thinking skills and to investigate whether relational thinking is a separable cognitive skill from the canonical EFs that contribute unique variance to students' math achievement. We assessed a large sample of elementary and middle school students on a newly developed tablet-based assessment battery that included ten standard tests of EF, a relational thinking task, a math fluency task, and a fraction comparison task.

Our relational thinking task contained two levels: children needed to determine whether the two pairs of shapes matched along a single dimension (Level 1) or two dimensions (Level 2). Performance on both task levels captured developmental improvements in relational thinking skills throughout elementary and middle school. We also found that performance on Level 1 of the task was significantly correlated with EFs, particularly working memory and task switching. However, relational thinking was a significant predictor of math achievement after accounting for the variance explained by EFs. This was true both for math fluency—which tests student's speeded ability to solve arithmetic, multiplication, and division problems—and fraction comparison. Furthermore, relational thinking was a significant predictor of fraction comparison performance even when math fluency, a domain-specific measure, was taken into account. These findings suggest that relational thinking is a distinct cognitive process that supports math performance in school-aged children over and above canonical EFs.

In prior work, it has been difficult to disentangle the influence of EFs and relational thinking because many standard assessments of reasoning are relatively complex and tax multiple abilities at once. However, the relational thinking task used here—particularly Level 1—was designed to minimize demands on EFs and other cognitive skills. By explicitly stating the different rule types prior to starting the game and by providing only two answer choices for each trial, Level 1 of the task focuses on children's ability to abstract a common relation between exemplars. Level 2, in which participants needed to identify matches in two dimensions rather than one, required participants to resolve increased relational complexity (Halford et al., 1998) and represent hierarchical rule structures (Bunge & Zelazo, 2006). However, it also presumably increased demands on canonical EFs, as participants had to override the previously learned rules of Level 1 (which was always explained and performed first). Critically, only data from the cleaner measure of relational thinking, Level 1, were included in the analyses examining relations with EFs and with math achievement.

To address the question of whether relational thinking is a distinct ability from the canonical EFs, we assessed which EF scores predicted performance on Level 1 of the relational thinking task, as well as how these scores related to children's math fluency and fraction comprehension performance. Consistent with previous studies that have documented relations between reasoning tasks that tap relational thinking and EFs (e.g., Fry & Hale, 2000; Richland et al., 2006; Richland & Burchinal, 2013; Starr et al., 2018; Thibaut et al., 2010; Thibaut & French, 2016), relational thinking scores were significantly correlated with EF scores—however, which EF components it correlated with varied as a function of age. Notably, the correlation coefficients for the associations between relational thinking and the canonical EFs tended to be even lower

than the correlations among the canonical EFs themselves. This result, together with the fact that no single EF factor was a consistent predictor of relational thinking across all age groups, suggests that relational thinking is separable from each of these other EFs.

Because EFs are conceptualized as domain-general cognitive processes that support academic performance, examining whether relational thinking predicts mathematical achievement is a criterion by which to assess whether it should be considered an EF. Indeed, we found that relational thinking and EFs each predicted unique variance in students' math achievement. The connection between EFs and math achievement is well-documented (see Bull & Lee, 2014 and Cragg & Gilmore, 2014 for reviews). However, individual differences in EFs do not explain all, or even a majority, of the variance in math scores. Here, we found that relational thinking predicted additional variance in math fluency and fraction comparison scores that was not accounted for by canonical EFs. Furthermore, in the case of fraction comparison, relational thinking predicted additional unique variance after accounting for both EFs and math fluency, meaning that the model already contained both domain-general and domain-specific predictors before we added in relational thinking. In fact, relational thinking was consistently the strongest domain-general predictor of math performance, judging from the pairwise correlations and linear regression model coefficients. Therefore, individual differences in students' relational thinking ability are predictive of achievement across multiple types of mathematical thinking.

Given the inherently relational nature of many mathematical concepts, it is not surprising that relational thinking contributes to math performance throughout grade school. Our findings are consistent with previous work demonstrating that reasoning relates to math achievement (Fuchs et al., 2006; Green et al., 2017; Taub et al., 2008), and suggest, specifically, that the relational thinking component of reasoning supports mathematical thinking. In addition, our findings add to the growing body of literature that suggests that relational thinking is particularly important for mathematical concepts like fractions and decimals (DeWolf et al., 2015; Kalra et al., 2020). For example, Kalra and colleagues (2020) found that relational thinking, as assessed by the Test of Relational Reasoning Jr. (TORR Jr; Jablansky et al., 2017) predicted fraction knowledge scores in 2nd and 5th graders even when controlling for a variety of domain-general (e.g., working memory) and domain-specific (e.g., math fluency) predictors.

Fractions are typically students' first exposure to number concepts beyond the natural numbers, and frequently they represent a stumbling block in math curricula (Siegler et al., 2013). In comparison to the natural numbers, fractions' bipartite structure (a/b) increases their relational complexity because students must process the value of each individual component as well as the overall value of the fraction. Instructional techniques that make the relational nature of fractions explicit (i.e., that fractions represent a relation between two numbers) may therefore help students make the conceptual jump from understanding natural numbers to understanding rational numbers (DeWolf et al., 2015). Indeed, several studies have demonstrated that pedagogical methods that explicitly encourage the use of relational thinking can improve students' ability to learn mathematical concepts (T. P. Carpenter et al., 1996; Kidd et al., 2008; McNeil & Alibali, 2005; Richland et al., 2004, 2012). An important future direction will be to

explore how increasing emphasis on relational thinking skills in math classrooms may improve student outcomes (Vendetti et al., 2015).

There is no single criterion of what makes a cognitive ability an EF. Most standard definitions reflect the idea that EFs are mid-level, domain-general, effortful cognitive processes that contribute to goal-directed behavior and support academic achievement. Much of the developmental psychology literature has focused on the three core abilities of inhibition, working memory, and cognitive flexibility (Diamond, 2013; Lehto et al., 2003; Miyake et al., 2000), but these three components are not necessarily exclusive. Furthermore, total independence has never been used as a criterion for considering two putative processes as distinct EFs; in fact, these three canonical EFs have been theorized to support one another (Diamond, 2013). The present data provide evidence that relational thinking is only weakly correlated with the canonical EFs and that it is independently related to academic achievement, as measured by two math tests. Therefore, consistent with previous views that relational thinking is central to human cognition (Alexander, 2016; Cattell, 1987a; Halford et al., 2010), we argue that relational thinking should be considered among the pantheon of EFs.

Our claim is based on analyses of data collected from a large, diverse sample of children in middle childhood who performed the ACE battery of cognitive tasks (Younger et al., 2022). However, this study is not without limitations. The ACE battery contains ten different cognitive tasks, nine of which were grouped through exploratory factor analysis into three EF composites. Relational thinking, on the other hand, was—along with task switching—measured using a single task. Because completion of the full ACE battery and scholastic assessments was already a multi-day endeavor, inclusion of additional tasks was not feasible. Importantly, however, the relational thinking task proved sensitive to capturing both developmental improvements and individual differences. An important future direction will be to assess relational thinking with multiple measures (e.g., the Test of Relational Reasoning-Junior, Jablansky et al., 2017). In addition, future work exploring the relation between EFs and academic achievement should also assess relational thinking to provide a more comprehensive view of the contributions of domain-general cognitive abilities.

In conclusion, the present work introduces a task that can be used to effectively measure individual differences in relational thinking ability throughout middle childhood. This task specifically focuses on the ability to identify and integrate abstract relations, while minimizing the demands on EFs. Individual differences in relational thinking predicted significant variance in students' math fluency and fraction comparison scores throughout middle childhood, even when variance from other EFs was accounted for. These results support our claim that relational thinking should be considered alongside the canonical EFs as a distinct core cognitive ability that uniquely contributes to academic achievement.

Chapter 3. Spontaneous and strategic relational offloading to physical space

3.1 General Introduction

One strategy for overcoming cognitive capacity limits during reasoning is to offload relational demand to external spatial representations. Despite a non-trivial literature on the cognitive science of visuospatial tools, research has focused mostly on implications for design and instruction (e.g., Bauer & Johnson-Laird, 1993; Franconeri et al., 2021; Hegarty, 2011; Shah et al., 2005; Shah & Hoeffner, 2002; Tversky, 2011), and few researchers have investigated these tools from the perspective of relational reasoning, complexity, and relational offloading. In particular, there are open questions about how and when individuals decide to strategically use visuospatial tools for offloading, how relational complexity changes when using an external tool, whether and how frequently individuals invent ad hoc tools to solve novel problems, and what cognitive processes led to the creation and adoption of the most widely used tools. Addressing these questions is critical because some researchers have theorized that there is a strong relationship between space and relational reasoning, which is best summarized by Dedre Gentner (2014) who called space “the universal donor of relational thinking,” but more work is needed to understand the mechanisms through which space supports relational reasoning.

The following section begins addressing these gaps in the literature by investigating the spontaneous and strategic recruitment of physical space to support problem solving. The study tests two hypotheses about relational offloading: either it is a specific strategy that is culturally transmitted—such as via learning and experience with writing systems, calendars, and other formal tools—or it is broadly available as part of our cognitive toolkit, separate from experience with these formal tools, and individuals can innovate ad hoc offloading strategies even in novel contexts. We worked with the Tsimane’, an indigenous farmer-forager people from the Amazon basin of Bolivia who live in a non-industrialized society and often have minimal levels of formal education and literacy. In one experimental condition participants needed to remember a sequential relation to answer the memory test questions and in the other condition participants needed to remember a categorical relation (i.e., preference between two options) to answer the questions. The findings provide insights about the origins of many visuospatial tools, such as graphs and diagrams, and suggest that innovating ad hoc spatial tools to offload relational demand is commonplace in everyday problem-solving.

3.2 Indigenous Amazonians spontaneously use space to offload cognitive demands

Abstract

Across many cultures, people use visuospatial tools to reduce the cognitive demands of thinking, reasoning, and remembering. However, little is known about the cognitive origins of these tools and the general strategy of recruiting space to solve novel problems. Here, we tested whether individuals from a non-industrialized society spontaneously and strategically used physical space to offload cognitive demands on a novel paradigm. We worked with the Tsimane', an indigenous farmer-forager group in Bolivia. Members of this group do not typically read, write, or draw, and often have little formal education. We found that children and adults ($N = 107$) spontaneously used spatial strategies to support their performance on the task, and, critically, that these strategies differed as a function of task goals. Further, participants increasingly used spatial strategies over the course of the experiment, even without feedback on performance. These results provide evidence that diverse human groups innovate ad hoc spatial tools to support cognition, complementing and extending prior studies that document such abilities in children and adults who live in societies where these skills are explicitly taught in formal educational settings.

Introduction

Humans have invented many spatial tools for reducing the cognitive demands of day-to-day tasks. For example, in high-literacy industrialized societies we regularly write to-do lists, draw diagrams, use calendars, and graph data (Gilbert et al., 2023; Hegarty, 2011; Tversky, 2015). These tools allow the user to offload demands onto physical space (Clark & Chalmers, 1998; Hegarty, 2010; Ishikawa & Newcombe, 2021; Risko & Gilbert, 2016), freeing up cognitive resources to think more abstractly (Atit, Uttal, et al., 2020; Gattis & Holyoak, 1996; Kirsh, 2010; Tversky, 2011), reason and remember more effectively (Brich et al., 2019; Gilbert et al., 2023; Kirsh, 2010; L. M. Padilla et al., 2018; Tversky, 2005), and even make discoveries (Tufte, 1983, 1997; Valleriani et al., 2023). However, despite their prevalence, little is known about the cognitive origins of these spatial tools and the general strategy of recruiting physical space to support solving novel problems. One hypothesis is that spatial strategies are rarely invented and are primarily learned through cultural transmission (B. Thompson et al., 2022). In this case, we would expect that spatial strategies for cognitive offloading would not be widely available to individuals unless they had been learned through sociocultural practices, such as formal schooling in which children learn specific spatial tools like drawing and writing with pen and paper. Alternatively, offloading to space may be a fundamental cognitive skill that is found across cultures, regardless of sociocultural practices. In this second case, we would expect individuals to innovate ad hoc spatial strategies for offloading even in novel contexts (Kirsh, 1995). Here, we test for the spontaneous use of space for cognitive offloading in a cultural context with little formal schooling and minimal engagement with reading, writing, and drawing.

Prior work on cognitive offloading has shown that between ages 6 and 10, children in WEIRD societies (Western, Educated, Industrialized, Rich, and Democratic; Henrich et al., 2010) begin to reliably and spontaneously offload demands onto external tools to facilitate problem solving,

such as physically rotating an object to reduce mental rotation demand (Armitage et al., 2020, 2022) or marking the hidden location of items with a pen to aid in retrieval (Bulley et al., 2020). However, research in industrialized societies alone cannot disentangle whether these strategies were learned through cultural transmission or discovered independently because there is widespread cultural support for formal spatial tools and strategies from a young age, such as in baby toys, books, and media (Armitage & Redshaw, 2022; Verdine et al., 2014). Though there has been prior work on spatial representations in non-industrialized societies, it has largely focused on the structure of internal representations of time and number (Boroditsky & Gaby, 2010; Dehaene et al., 2008; Fedden & Boroditsky, 2012; Floyd, 2016; Le Guen & Balam, 2012; R. Núñez, Cooperrider, Doan, et al., 2012; R. Núñez, Cooperrider, & Wassmann, 2012; R. E. Núñez & Sweetser, 2006; Pitt et al., 2021). Some relevant work has investigated the origins of linear order, a specific spatial organization practice (Cooperrider et al., 2017). Unschooling Yupno adults spontaneously arranged objects in linear orders based on size and numerosity, but with less consistency than US adults (Cooperrider et al., 2017). By contrast, the present study investigates the spontaneous and strategic recruitment of physical space for supporting problem solving and examines whether individuals use different spatial strategies as a function of task goals.

We worked with the Tsimane', an indigenous farmer-forager people from the Amazon basin of Bolivia. The Tsimane' live in a non-industrialized society with often minimal levels of formal education and literacy (for details on the Tsimane' cultural context, see Huanca, 2008 and O'Shaughnessy et al., 2023). We recruited both children and adult participants to capture potential developmental differences as well as differences in daily experiences due to age. In their day-to-day activities, the individuals we worked with do not typically read, write, draw or use paper, linear measurement tools, watches, calendars, maps, phones, or other formal spatial tools, which influence spatial conventions and offloading strategies (Bergen & Lau, 2012; Cooperrider et al., 2017; Fuhrman & Boroditsky, 2010; Gilbert et al., 2023; Grinschgl & Neubauer, 2022; Pitt et al., 2021; Risko & Dunn, 2015; Starr & Srinivasan, 2021; Uttal, 2000). Thus, testing the spontaneous use of space in this group shows whether the general strategy of spatial offloading is shared broadly among human groups with markedly distinct cultural experiences.

Experimental paradigm

We designed a novel paradigm that taxed participants' working memory using sets of laminated cards with faces on them. Participants were handed eight cards, one at a time, and told a piece of information about each individual that they were asked to remember for a subsequent memory test. In the Order condition, participants were told the order in which the individuals arrived at a market (Figure 3.2.1A). In the Preference condition, participants were told the individuals' preferences (e.g., plantains vs. coconuts) (Figure 3.2.1A). At the end of each trial (Figure 3.2.1B), participants were asked two questions, either about the sequence in the Order

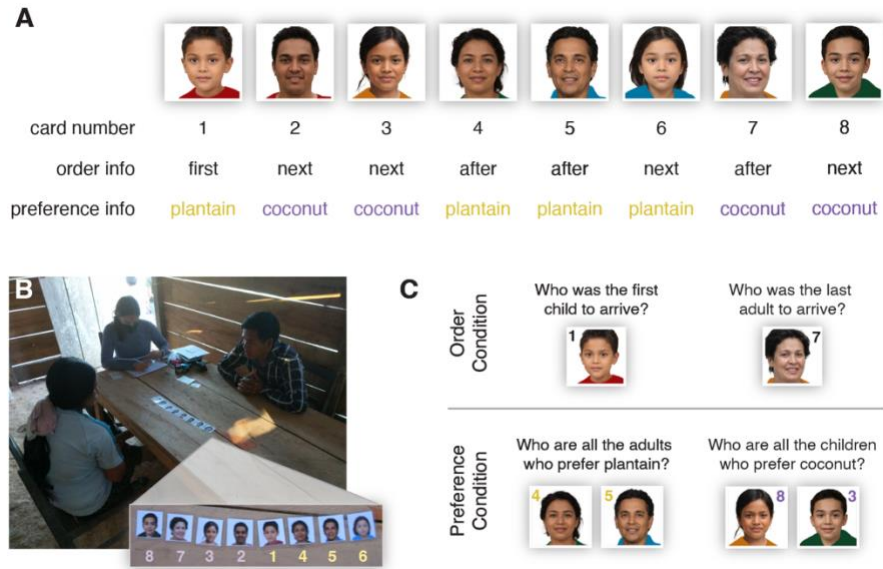


Figure 3.2.1. Sample stimuli and condition-specific prompts for the memory task. (A) Face card stimuli for the first trial. Numbers indicate the order in which cards were handed to participants in all three conditions, which was also the order they “arrived to market” in the Order condition. Participants were told either the order or preference information when each card was handed to them, depending on the condition. For example, for the first card participants heard either “First this boy arrived” (Order condition) or “This boy prefers to eat plantain” (Preference condition). In the Control condition, no information was shared about the cards as they were handed to participants. Face images were adapted from [Generated.Photos](#). (B) An adult in the Preference condition participating in the study on the first trial. The inlaid image shows a closeup of her final card layout. The number annotations show the order that the cards were distributed, and the colors represent the preferences (yellow for plantains and purple for coconuts). This participant organized the cards by preference (grouped organization), creating a line shape. (C) The memory questions asked on the first trial for the Order and Preference conditions with the correct answers.

condition or about the preferences in the Preference condition (Figure 3.2.1C; Materials and Methods). Participants completed four of these trials and condition was manipulated between participants.

One strategy to facilitate memory retrieval afforded by this paradigm was to place the cards on the table in a spatial arrangement that represented the to-be-remembered information, for example, by sequentially ordering the cards in the Order condition or grouping the cards by preference in the Preference condition. Critically, participants were never instructed on how to place cards or to use these strategies. Additionally, because card games are not prevalent in Tsimane’ culture, it is unlikely that participants had prior experience with spatially arranging cards (e.g., ordering, sorting, etc.). We were particularly interested in the first trial (T1), which captured participants’ spontaneous use of space during their first experience with this task, even before they had heard any memory questions. If the use of space for offloading is universal, we would expect participants to spatially arrange the cards in a manner that supports memory retrieval beginning on T1. Additionally, this design allowed us to detect spontaneous changes in strategy across the four trials because participants did not receive feedback on their response accuracy. We also included a Control condition in which participants were instructed to place cards on the table in any way they wanted to examine card placement without a memory demand.

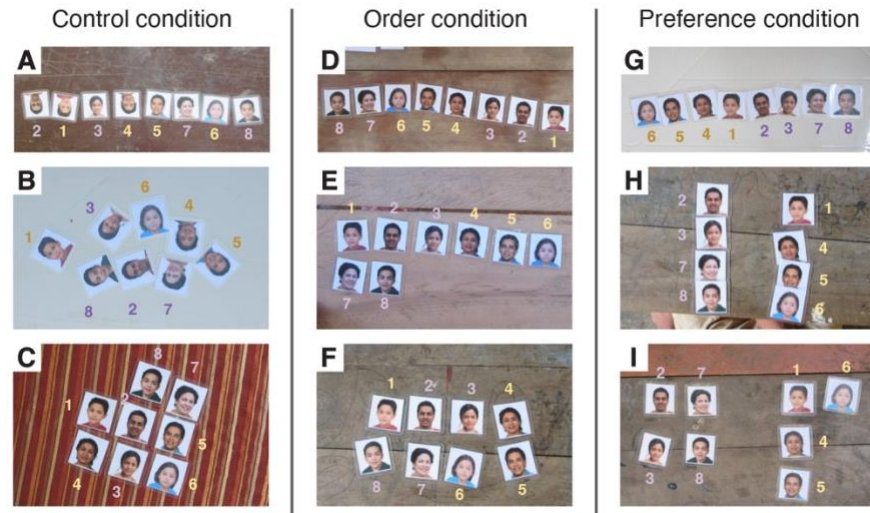


Figure 3.2.2. Annotated images of card layouts. Resulting card layouts from nine different participants on the first trial by condition. Number annotations show the order that the cards were distributed, and colors represent preferences (yellow for plantains and purple for coconuts). Images were coded for shape (visual form of the layout) and organization (information represented by the card placement: sequential order, grouped by preference, or neither). An organization was coded as “sequentially ordered” if the layout preserved the sequence that the cards were distributed. An organization was coded as “grouped” if a straight line could be drawn through the layout that separated the cards belonging to each preference. Order and Preference condition images show the most common shapes created when representing the condition-relevant information (sequential order and preference groups, respectively). Control condition images show the three most common shapes. These layouts demonstrate the variability in shapes within and between conditions. See Figure 3.2.S6 for original uncropped and unannotated images and Movie S1 for all images from the first trial.

We coded two features of the resulting card layouts: shape and organization (Materials and Methods). Shape refers to the overall form of the card layout, such as a line (Figure 3.2.2A, D, G), rectangle (Figure 3.2.2F), or two clusters (Figure 3.2.2H, I). Organization refers to what information, if any, is represented in how the cards were placed in relation to one another: either sequentially ordered (Figure 3.2.2D-F), grouped by preference (Figure 3.2.2G-I), or neither (Figure 3.2.2A-C). If participants spontaneously use space for offloading, we predicted that the organization of cards should differ by condition on T1. Specifically, the cards should represent sequential order in the Order condition and preferences in the Preference condition. We were also interested in the shape of card arrangements when they were organized in these strategic ways. Because there are few practices in Tsimane’ culture that would establish visuospatial conventions, we predicted variability in shapes between individuals within a condition. We also report performance across all four trials and highlight two participants as case studies to examine changes in spatial strategy use.

Participants spatially organized cards to strategically represent relevant information

For both children and adults, Fisher’s Exact Test revealed a significant effect of condition (Order, Preference, and Control) on organization (sequential, grouped, or neither) on T1 ($ps < .001$; Figure 3.2.3A). Children and adults in the Order condition organized their cards in sequential order significantly more often than in the Control and Preference conditions (Figure 3.2.3A, Table 3.2.S4). This result suggests that even though the cards were given in the same order in all three

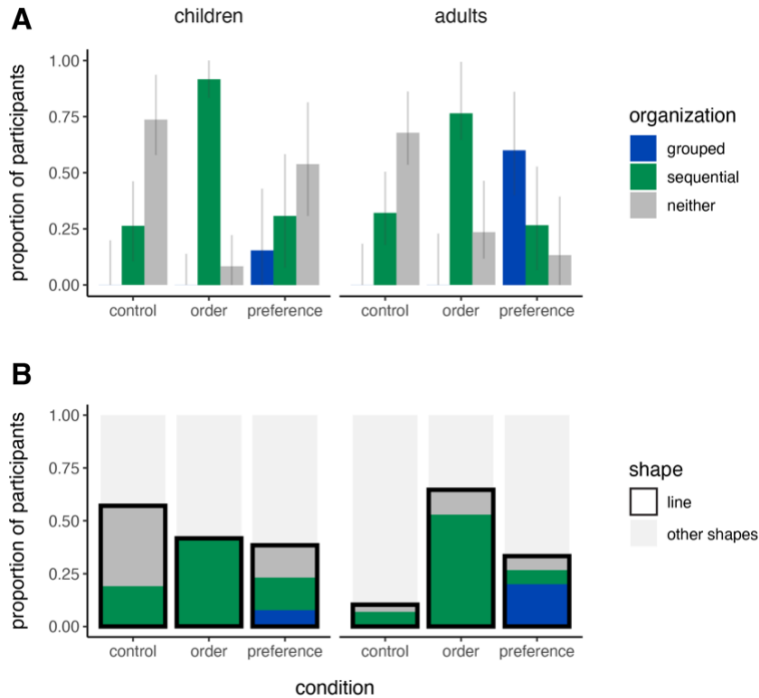


Figure 3.2.3. Organization and shape by memory condition. (A) Proportion of participants on the first trial who represented preference groups, sequential order, or neither in their card organization. Error bars show the multinomial 95% confidence interval. (B) Proportion of participants in each condition who used a line shape (black box) versus other shapes (light gray) on the first trial. The participants who created lines are further subdivided into the organization represented by the line: grouped, sequential, or neither. Note that organization for the other shapes is not shown.

conditions, it was not generally the default for participants to organize cards sequentially—neither when there was no task demand (Control) nor when the task did not require memory of the sequence (Preference). Rather, it suggests that participants in the Order condition strategically organized cards to alleviate the need to remember the sequential order.

While the Order condition instruction to “remember the order” explicitly hinted at what organization would be helpful, the Preference condition instruction to “remember the preferences” gave no such hints, making no mention of “groups” or that organizing by preference type would be useful. Thus, it is notable that on T1 of the Preference condition 9/15 adults organized their cards into groups by preference, and even two children used this spatial strategy (Figure 3.2.3A), whereas no participants in the other conditions happened to organize cards this way without preference information (see Supplementary Materials for full results and discussion of age effects). These results are consistent with the prediction that participants would strategically position cards with the same preference close to each other to reduce the number of individual preferences to remember.

Further evidence comes from examining the relation between card organization and accuracy on the two memory questions. To answer these questions correctly, participants needed to integrate the to-be-remembered information from the entire set of eight cards (Figure 3.2.1C). If offloading to space is a beneficial strategy, then the participants who spatially represented this information

should be more accurate on the test questions. This was clear on T1 of the Preference condition, in which participants who grouped by preference were more accurate on the test questions ($M = 1.09$ out of 2) than those who did not group ($M = 0.06$; $t(11.11) = 4.01$, $p = .002$). Notably, across all trials the odds of answering a question correctly were 31.54 times greater when the cards were grouped versus not grouped ($p < .001$; Table 3.2.S5). Testing this prediction within the Order condition was more difficult because of the near-ceiling effect of participants sequentially ordering their cards on T1 (24/29) (see Supplementary Materials).

Given that many participants did report some schooling (Materials and Methods), it is central to the interpretation of the results to confirm that strategically organizing the cards on T1 was not simply due to schooling or literacy. Indeed, neither years of school ($b = 0.07$, $p = .727$) nor literacy ($b = 0.21$, $p = .112$) predicted organizing by preference on T1 of the Preference condition. Of the 11 participants who grouped, four were not literate (3 adults and 1 child): one adult who reported no schooling (Participant MP in *Brief Case Studies*, Figure 3.2.5) and the rest reported four or fewer years. In the Order condition, there was not enough variation in organization to test for these effects, but among the 24 participants who sequentially ordered on T1, nine were not literate (2 adults and 7 children). The two non-literate adults reported no schooling (Figure 3.2.S5), and the seven children reported six or fewer years. For additional details on these analyses, see Supplementary Materials. Together, these results show that this strategy is available even in the absence of schooling and literacy.

Shape of card layouts varied within and between conditions

The resulting shapes of strategically organized card arrangements varied both within and between conditions (Figure 3.2.2, Figure 3.2.S3). In the Order condition, six different shapes were created when representing sequential order; line (58.3%; Figure 3.2.2D), line + extra (16.7%; Figure 3.2.2E), and rectangle (12.5%; Figure 3.2.2F) were the most common, and the other three shapes each appeared just once. In the Preference condition, five different shapes were created when organizing by preference; line (36.4%; Figure 3.2.2G) and two clusters (36.4%; Figure 3.2.2H, I) were the most common, and the other three shapes each appeared just once. In the Control condition, when there was no memory prompt and therefore no motivating organization, nine different shapes were created; line (30.0%; Figure 3.2.2A), random/unknown (18.0%; Figure 3.2.2B), square (16.0%; Figure 3.2.2C), and rectangle (16.0%) were the four most common configurations. If participants had been influenced by culturally predetermined spatial conventions and strategies, we would have observed more convergent use of space (Cooperrider et al., 2017; Pitt et al., 2021; Starr & Srinivasan, 2021). However, we observed substantial within-condition variability in shape (Figures 3.2.S2, 3.2.S3), providing evidence that offloading to space was an ad hoc strategy innovated by individual participants in the context of this task.

We also found significant differences in the distribution of shapes between conditions ($p = .006$; see Supplementary Materials), though lines were the most common shape across conditions. Perhaps surprisingly, the creation of a line did not always indicate that the participant was representing sequential information. Indeed, some participants created lines when representing grouped information (e.g., Figure 3.2.2G), and still others created lines that represented neither sequential nor grouped information (e.g., Figure 3.2.2A and Figure 3.2.5, middle row, trial 1).

Critically, lines were used more frequently and strategically in the Order condition, with some age group differences (Figure 3.2.3B). Specifically, the odds of an adult creating a line were 14.63 times greater in the Order condition versus the Control condition ($p < .001$), suggesting that adults in the Order condition were specifically selecting lines to represent sequential order and not simply placing cards in a line as they received them (Figure 3.2.3B). Adults also created more sequentially ordered lines in the Order condition than in the Preference condition ($OR = 14.00, p = .036$). Children, on the other hand, created lines equally often across the three conditions ($p = .599$), but notably all the lines created in the Order condition were sequentially ordered, whereas only a third were sequentially ordered in the Control condition (Figure 3.2.3B; see Supplementary Materials for additional analysis details). Thus, although lines were the most common shape, both children and adults predominantly created sequentially ordered lines in the Order condition.

Use of spatial strategies increased after the first trial

Although T1 was of greatest interest, we also included three additional trials in the Order and Preference conditions to assess whether participants would use a spatial strategy at any point during the task. In the Preference condition, the odds of children using a spatial grouping strategy increased after T1 ($b = 1.71, p = .027$; Figure 3.2.4). For adults, more than half grouped by preference on T1, and though the effect of trial did not reach significance ($b = 0.83, p = .058$), including trial as a predictor explained significantly more variance than the model without it ($\chi^2(1) = 4.56, p = .033$). See Supplementary Materials for full analysis details as well as Order condition results (Figure 3.2.S4).

The increased incidence of grouping on subsequent trials was accompanied by qualitative changes in the shapes created. When the cards were grouped by preference, sometimes a gap was visible between the groups of cards (e.g., Figure 3.2.2H, I). Notably, the proportion of participants who included a gap increased between T1 (4/11) and T4 (12/20). We even observed

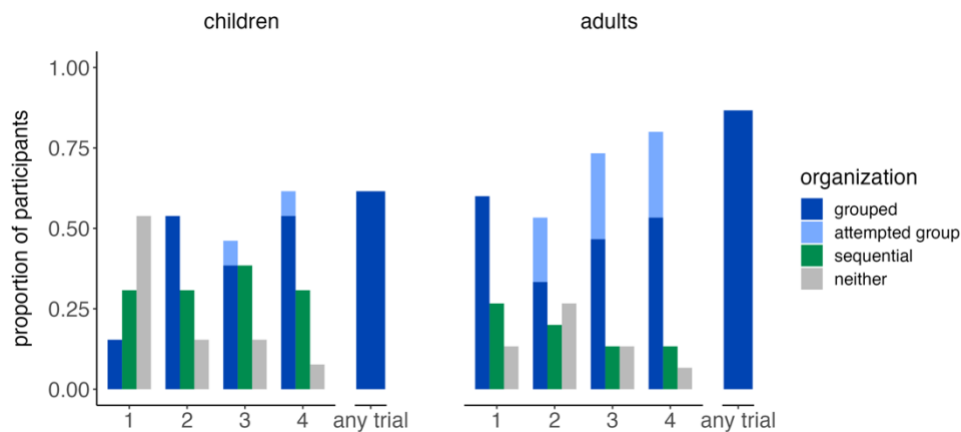


Figure 3.2.4. Spatial organization across all trials in Preference condition. Proportion of participants on each trial of the Preference condition who used each organization: grouped, attempted group, sequentially ordered, or neither. The last bar represents the proportion of participants who correctly grouped on at least one of the four trials. “Attempted group” refers to the case where participants were using the spatial grouping strategy but the resulting layout was incorrectly grouped due to an error when placing one or two cards (e.g., Figure 3.2.5, top row, trial 3; Materials and Methods).

a new two-cluster shape after T1—a line with a gap in the middle separating the two halves (e.g., Figure 3.2.5, middle row, trial 4)—which was made by three participants who had each started the task creating unbroken lines. Compellingly, this new shape did not appear on any trial in the other conditions. Additionally, accuracy on the memory questions increased over the trials (Table 3.2.S5). The increase in the use of space—both in the organization and shape of layouts—provides evidence of strategy change towards more memory-efficient external representations. Importantly, because performance feedback was not given, strategy changes were likely due to metacognition about how best to solve the problem (Dunn & Risko, 2016).

Case studies show change in strategy

Although participants reported little formal schooling (children: $M = 4.54y$, $SD = 1.96$, $range = [1, 9]$; adults: $M = 4.36y$, $SD = 3.58$, $range = [0, 13]$), the responses of those with no schooling are especially informative about people's spontaneous use of space. In this section, we focus on two of the three adult participants in the Preference condition who reported no schooling—MP and AS (see Supplementary Materials for full descriptions, including for third adult CM, videos, and Order condition cases, Figure 3.2.S5). These participants were not literate in Tsimane' or in Spanish. Interestingly, all three participants grouped the cards by preference on at least one trial. However, each participant engaged with the task in notably different ways.

MP was approximately 40 y of age. On T1, she grouped the cards by preference and answered both memory questions correctly. She continued this accurate performance, except on T3, where she attempted to group but made an error placing the cards. From a visual scan of her annotated trial images (Figure 3.2.5, top row), it is apparent that as the trials progressed, the gap between the two groups increased. This suggests that she was leveraging the affordance of space to facilitate her accurate performance, both to offload group membership (grouped organization) and the boundary between the groups (shape: increased spatial separation). A second interesting feature of MP's trials was the way she placed the cards, starting from the middle of the table, and moving outward to the left and right. This direction differs from many WEIRD populations—which tend to prefer the direction of their written language, such as left-to-right (Bergen & Lau, 2012; Cooperrider et al., 2017; Fuhrman & Boroditsky, 2010). Moreover, even participants with some formal schooling used this “center-out” strategy (e.g., Figure 3.2.2G), suggesting that this behavior reflects a strategy to solve this novel task rather than, for example, spatial biases created by writing systems.

AS was 58 y of age. Unlike MP, AS did not spontaneously group cards by preference on T1 (Figure 3.2.5, middle row), and she did not answer any of the memory questions correctly on the first two trials. On T3, she again created a line but began attempting to organize the cards by preference. Although neither the resulting grouping nor the answers to the memory questions were correct, her intention to represent the two preference groups in her spatial organization was clear from her response behavior: she pointed to the five cards on the right to answer one preference and to the three cards on the left to answer the other preference (Movie S3). On the very next trial, not only did she correctly group the cards by preference (Figure 3.2.5, middle row, trial 4), but strikingly, she placed the cards with a gap between the two groups, visually separating

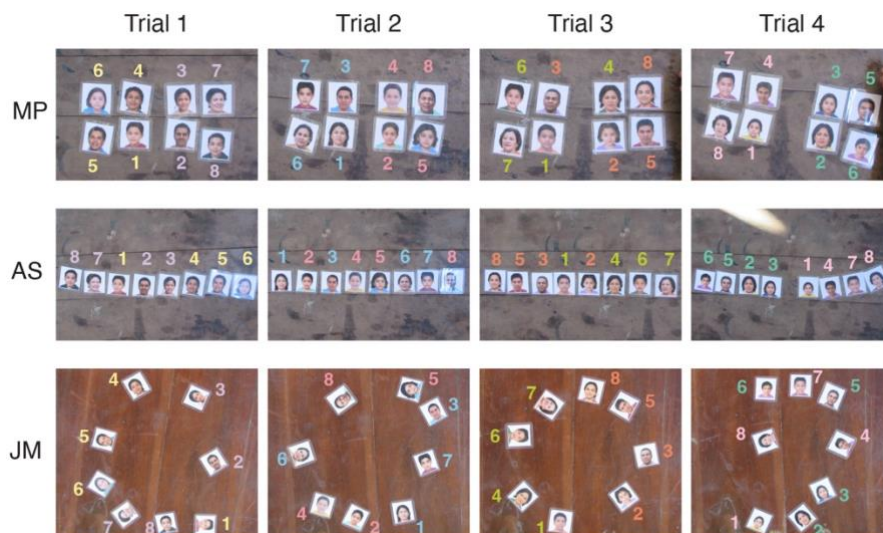


Figure 3.2.5. Annotated trial images of case study participants from the Preference condition. Trial images for the three participants in the Preference condition who reported zero years of schooling. The number annotations show the order that the cards were distributed, and the colors represent the preferences. See Figure 3.2.S1 for list of preferences and 3.2.S7 for original uncropped and unannotated images.

them. Further, she got one memory question correct and the other one partially correct (i.e., indicating all participants with a given preference, rather than only the females with that preference). We interpret this progression as evidence that AS was discovering and fine-tuning an ad hoc spatial strategy—even without feedback. Together, these case studies demonstrate how spatial strategies are deployed, and possibly invented, on a novel task by individuals who have no experience with formal visuospatial tools or schooling.

Discussion

Our findings show that many Tsimane’ children and adults spontaneously represent relevant information in space to reduce cognitive load, in turn facilitating memory performance. They use this strategy even in the absence of any instruction on how space could be useful and with minimal or no formal schooling and literacy. Participants used space in a variety of ways: some created the same shapes with the card stimuli (e.g., lines) but to represent different information (e.g., orders or groups), whereas others represented the same information but with different shapes. Such variability suggests that strategies are created ad hoc by individuals during the task rather than determined by prior spatial conventions. This conclusion was further supported by changes in strategies across trials (Figure 3.2.4), which happened without feedback on performance and were particularly clear in the case studies of participants in the Preference condition who reported no formal schooling (Figure 3.2.5). These changes involved spatially organizing the cards by preference to offload group membership and increasing the spatial separation between the groups to offload the boundary between them.

These findings complement prior studies on cognitive offloading (Armitage et al., 2020, 2022; Armitage & Redshaw, 2022; Bulley et al., 2020; Dunn & Risko, 2016; Gilbert et al., 2023; Risko & Gilbert, 2016; Verdine et al., 2014) by adding results from a non-WEIRD population, a known gap

in this line of research (Armitage & Redshaw, 2022). Working with the Tsimane' allowed us to disentangle the presence of offloading strategies from routinized exposure to external thinking tools transmitted via formal schooling, which may ingrain this style of thinking from a young age. Though it is possible that simple exposure to any amount of schooling could be sufficient to promote the strategic use of space on this task, the results from participants who were not literate and reported no schooling provide evidence that schooling is not required to use space in this way. Further, the Tsimane' cultural context, specifically the representational practices and visual culture, are vastly different from the cultural contexts in WEIRD societies. For example, most participants are not seeing or engaging with written text on a daily basis, even those who are literate. Therefore, the findings of this study provide insight into the recruitment of space as a tool in a different cultural context and support the hypothesis that cognitive offloading to space is a fundamental human capacity. By extension, these results suggest that innovating ad hoc offloading strategies is available in our day-to-day lives as part of our problem-solving cognitive toolkit, regardless of our experience with formal visuospatial tools.

Our findings also support theories that visual representations of sequential order and spatial grouping are fundamental spatial organization principles (Cooperrider et al., 2017; Ingold, 2007; Kirsh, 1995; Tversky, 2011), even when they are less present in an individual's surrounding visual culture. Extending these theories, we found that participants strategically selected a spatial organization as a function of task goals. Dovetailing with prior work (Cooperrider et al., 2017; Pitt et al., 2021; Starr & Srinivasan, 2021; Tversky et al., 1991), we also found that without the conventionalized spatial practices transmitted via formal schooling, how space is recruited to support thinking remains flexible and can vary between individuals.

We conclude that diverse human groups share space as a medium for offloading cognitive demands, but that the specifics of how space is recruited to support thinking can vary between individuals. This human capacity paired with between-individual and cultural variability may in turn explain why visuospatial representations are used across many cultures—writing systems, calendars, number systems, graphical representations, etc.—but the precise forms of these representations are diverse and varied.

Materials and Methods

Participants

A total of 46 children ($< 16y$, $n_{\text{female}} = 21$) and 63 adults ($\geq 16y$, $n_{\text{female}} = 30$) participated in the study from eight Tsimane' communities near San Borja, Bolivia. Participants provided informed consent and were compensated with goods for their participation. Two adult participants were removed from the data set before analysis: one because she was originally from a Bolivian city, and one because he had already participated in a different study that could have influenced his behavior in this study. Table 3.2.S1 shows age, schooling, and literacy for the 107 children and adults included in analysis. The study was performed in accordance with the Institutional Review Board at the University of California, Berkeley and with permission from the Gran Consejo Tsimane' (Tsimane' Grand Council). We received consent from all adult participants, and assent from all children participants with consent from their legal guardians. Informed consent and

assent were obtained in Tsimane' in an ethical way via existing, culturally appropriate explanations and forms.

The age cutoff of 16 years old between children and adults was used because it is around this age that individuals typically start having children in Tsimane' society. Further, individuals may attend school in some capacity until they are married or have children. Thus, this age cut off is not strictly developmental in nature, but also describes differences in experiences between individuals in different stages of life.

Note that there is a great deal of variability in the self-reported measure years of school because what constitutes a completed year of school can vary drastically between participants and is different than one year of school in an industrialized society context. For example, some participants may have attended school only once per week for 2 hours and reported that as one year of school. Further, what schooling entails differs depending on the individual's age and what community they are from, and not all participants who have attended school are literate. In our sample, 10 adults reported no schooling (16.95%), and 81.36% had no education beyond sixth grade, a benchmark that has been considered "unschooled" in previous research with non-industrialized groups (e.g., Cooperrider et al., 2017). 32.79% of adults were not literate, and 57.38% of adults scored the maximum points on the literacy test. Though all children reported at least one year of school, 86.96% had six years or less, and 58.70% were not literate. Only 8.70% of children scored the maximum points on the literacy test.

Procedure

Experienced translators who were fluent in Spanish and Tsimane' were provided by the Centro Boliviano de Investigación y de Desarrollo Socio Integral (CBIDSI). All participants first consented to participation then answered demographic questions (e.g., age and years of school). They also completed brief tasks to measure numeracy, Spanish fluency, and Spanish and Tsimane' literacy. Consent and the demographic survey were administered by one or both CBIDSI coordinators, one of whom speaks some Tsimane', and they were typically joined by one translator. After completing these intake tasks, the participant completed the memory task. Participants were randomly assigned to one of the two experimental conditions (Order or Preference). Random assignment was done within age group (children or adults) to ensure balanced sample sizes (Table 3.2.S2). The Control condition was run after data collection for the experimental conditions had been completed.

The two experimental conditions (Order and Preference) were administered by the experimenter (first author) in Spanish and translated into Tsimane' by an experienced translator. The experimenter always sat to the left of the participant, with the translator across the table or to the right. The Control condition was administered by the same research coordinator who administered the demographic intake survey with a translator, except for the first eight Control participants, who were run by the experimenter and a translator. The research coordinator did not consistently sit in the same position relative to the participant, but typically sat either across from or to the left of the participant.

Memory Task

The stimuli consisted of four sets of eight mixed age and gender faces printed on card stock, cut into 2.5 x 2.5 inch squares, and laminated (Figure 3.2.S1). There were two between-subject experimental conditions—Order and Preference—and a Control (no-task) condition. The general procedure for the two experimental conditions was identical. Participants first heard condition-specific instructions, including what information to remember (i.e., Order: “I’m going to tell you the order in which they arrive at the market, and I want you to remember this order.”; Preference: “I’m going to tell you about their preferences, and I want you to remember these preferences.”), then were handed the cards one by one and told the relevant information for each card (e.g., Order: “This boy arrived first.”; Preference: “This boy prefers plantains.”; Figure 3.2.1A). After the last card, the translator took a picture of the card layout on the table. Finally, participants were asked two questions to test their memory of the relevant information (e.g., Order: “Who was the first child to arrive?”; Preference: “Who were all the adults who preferred plantains?”; Figure 3.2.1C and Table 3.2.S3); they received no feedback on their answers. This procedure was repeated for a total of four trials. The cards were handed out in the same order across all three conditions.

The instructions for the Order and Preference conditions mentioned that participants could put cards on the table if they wanted to. This option was mentioned explicitly because the card layout was essential to our study, and piloting revealed that participants were unfamiliar with using or holding cards and were unsure whether they could use the table instead of holding and organizing them in their hands. If a participant did not put the cards on the table on the first trial, the translator prompted the participant to do so (e.g., participants were holding the cards in their hands and/or cards were falling on the ground). Critically, however, neither the translator nor experimenter stated how to use the table or how to place the cards. Some participants continued holding the cards, and the picture is of the cards in their hands.

The procedure for the Control condition was the same as for the experimental conditions except that participants were not instructed to remember any information, nor given any information as the cards were handed out. Instead, participants were instructed to put the cards on the table in any way that they wanted. Also, the Control condition had two trials instead of four.

This task was designed with extensive feedback on the procedure and instructions from the CBIDSI coordinators and translators to ensure that it was culturally appropriate and made sense to participants. See Supplementary Methods for additional details about this task and condition-specific instructions.

Data Coding. Card layouts were coded for three features—shape, organization, and directionality—from the photos of the table that were taken at the end of each trial. Video recordings were consulted where necessary to confirm coding.

The shape of each card layout was coded by two independent coders using a shape coding guide (<https://osf.io/75vrj/>). Interrater reliability was high (Cohen's $\kappa = .91$), and any disagreements were resolved by the first author. The two coders and the author did not know the

condition the participant had been assigned to or the organization of the layout when making shape judgments. Images were coded in a random order, shuffled across all trials and conditions.

After shape was coded, the organization of each card layout was coded by the first author as either “sequentially ordered,” “grouped,” or “neither.” A layout was coded as sequentially ordered if it preserved the sequence that the cards were handed out in, regardless of condition. A layout was coded as grouped if a straight line could be drawn (horizontal, vertical, or diagonal) that perfectly separated the cards belonging to each preference. This was coded independent of whether there was a visible gap between the two groups in the shape. For the Preference condition trials, we also coded for “attempted grouping,” which was when it was clear from the card placement and the participant’s responses to the memory test questions that they had been attempting to use a spatial grouping strategy, but made a small execution error (i.e., misplaced one card, or swapped two cards with each other), resulting in the final layout not being correctly grouped (e.g., Figure 3.2.5, top row, trial 3). Even with this additional code, the organization coding was conservative and likely underestimates the prevalence of the spatial grouping strategy because of the strict requirement for the grouping to be correct in order to be coded as grouped or one card away from correct to be coded as attempted. A layout was coded as neither if it did not meet the requirements for sequentially ordered or grouped. See Supplementary Methods for additional details about coding organization.

The directionality of the cards was also coded by the first author. This feature captured the direction that the cards were placed on the table relative to the participant. For example, cards could be placed left-to-right, right-to-left, top-to-bottom, bottom-to-top, or some combination if there were multiple rows or columns. Some layouts used what we called a “center out” pattern, in which participants started by placing cards in the relative center and building subsequent cards out from there, typically to the left and right. When the shape of the cards was “random” or “unknown”, the direction was coded as “none.”

Literacy Measure

To measure literacy, participants were asked to read four short, simple sentences, two in Spanish and two in Tsimane’ (e.g., “El gato tiene miedo.” [The cat is scared.]). The research coordinator rated their reading on a 3-point scale: 0 = “none”, 1 = “some”, and 2 = “perfect.” Note that the coordinator was assessing reading skill only, not comprehension. Literacy score is the sum score of the four questions (min = 0, max = 8).

Data Analysis

All analyses were run in R version 4.2.1 (R Core Team, 2022). See Supplementary Methods for the list of R packages used for analysis and visualization. For all primary analyses, children and adults were analyzed separately because of the possible difference in current exposure to formal schooling in day-to-day activities. Even though schooling is variable and relatively minimal, children (<16y) are likely to be attending school in some capacity based on their age and not having children of their own. This means that children may have more recent and immediate exposure to spatial tools (e.g., schoolbooks, numbers), whereas adults are likely to no longer

attend school, and therefore are not typically exposed to or using these tools on a day-to-day basis. For this reason, we kept the age groups separate for analysis unless otherwise noted.

To test for associations between condition (Order, Preference, Control) and the outcome measures of interest (organization and shape), we used Fisher's Exact Test. Though a χ^2 Test of Independence is the most common statistical test used, there were some cases in our data in which at least one of the expected value cells was less than 5, violating the assumptions of this test. Therefore, for consistency we use Fisher's Exact Test for all cases testing the association between two categorical variables, including follow-up pairwise comparisons between conditions. Fisher's Exact Test outputs only a p-value. When it is run on a 2x2 contingency table, an effect size—an odds ratio (OR)—and its 95% confidence interval can also be calculated from the contingency table and are reported. The OR gives the odds of an outcome occurring in one condition (e.g., sequentially ordering in the Order condition) relative to the odds of that outcome occurring in another condition (e.g., sequentially ordering in the Control condition). In follow-up pairwise comparisons, we used Bonferroni correction to correct for multiple comparisons, multiplying the p-value by 3 since we did three pair-wise tests (Order vs Control, Order vs Preference, and Preference vs Control). See Supplementary Methods for additional details about Fisher's Exact Test and its interpretation.

Models to test for effects of organization on memory test accuracy, effects of schooling and literacy on organization, and changes in spatial strategy use over trials were run separately for the two experimental conditions because of the near-ceiling effect of participants sequentially ordering their cards on T1 in the Order condition. We used two-tailed t-tests to test for differences in accuracy on the T1 memory test questions between participants who sequentially ordered versus not in the Order condition and between those who grouped versus not in the Preference condition. We also used logistic mixed effects models with a random intercept for participant to test that the effects of spatial organization on accuracy held across trials. These models also allowed us to test for increases in accuracy over the trials. All analyses with accuracy as the outcome variable collapsed across age groups. To test for effects of schooling, literacy, and age on T1 organization, we collapsed across age groups and used a logistic regression predicting grouping versus not on T1 of the Preference condition. We did not run this model for the Order condition because the high proportion of participants who sequentially ordered left little between-participant variation to be explained. However, we did look qualitatively at the schooling and literacy of participants who did and did not sequentially order on T1. Finally, to test for changes in spatial strategy use we used logistic mixed effects models predicting organization from a fixed effect of trial and a random intercept of participant. These models were run separately for each age group.

Data and Materials Availability

All data needed to evaluate the conclusions in the paper are present in the paper or the Supplementary Materials. Materials, data, and analysis files are available through Open Science Framework at <https://osf.io/75vrj/>. Additional data related to this paper may be requested from the authors.

Chapter 4. Scaffolding relational reasoning: A promising approach for promoting graph comprehension

4.1 Abstract

Knowing how to interpret graphs and make decisions based on the information presented are skills that are increasingly important in the workplace and in daily life. Acknowledging their importance, U.S. math and science standards include graph comprehension in late elementary and middle school. However, children—and even adults—struggle with these skills. One source of difficulty is with mapping the perceived visuospatial patterns to the real-world meaning of these relations. This mapping can be construed as a form of relational reasoning, which is the capacity to map multiple relations between representations. Here, I first propose that research on graph comprehension—from cognitive models to pedagogy—would be strengthened by considering relational reasoning as a foundational cognitive skill. Then, I report the results of a preliminary study that I designed and conducted based on this perspective. The study evaluates the benefit of emphasizing relational reasoning as part of a graph comprehension intervention for middle schoolers. I developed two well-matched lessons on y-intercept and slope, focusing students' attention either on the higher-order relations represented by graphs or on graphs' visual features. 289 U.S. public school students in grades 5-7 (ages 10-13) participated in this two-day study conducted remotely in their math class. The results suggest that both lessons were effective in improving students' knowledge of y-intercept and slope as well as reasoning with graphs in the transfer task. I conclude by discussing the potential benefits of promoting relational reasoning and visual pattern recognition in graph comprehension and point to many future research directions.

4.2 Introduction

In 1854, London doctor John Snow made a groundbreaking discovery. By showing that cholera was linked to contaminated drinking water, he proved that this deadly disease was waterborne as opposed to airborne, which was the prominent theory at the time. Dr. Snow's contribution to science was not only this discovery, but also the tool he created to uncover this elusive relationship. Searching for patterns in deaths due to cholera, Dr. Snow marked their locations on a map and noticed a concentration of deaths around a single water pump, which was later found to be polluted by sewage. This now-famous early example of data visualization allowed Dr. Snow to *see* abstract patterns in data and make inferences about the relation between water and cholera that were not otherwise apparent.

Data visualizations, such as Snow's cholera map, are designed to leverage our visual system's fluency perceiving visual patterns, using visual features such as spatial position, color, and area to represent information that may not typically have these properties (Bertin, 1983; Hegarty, 2011; Tversky, 2011). As the presence of data in our daily lives continues to increase, so too does the need to understand and visualize it (Börner et al., 2019; Franconeri et al., 2021; V. Lee & Wilkerson, 2018). Data-related jobs are an increasing sector of the job market (Bonesso et al., 2020; Gardiner et al., 2018), data skills are in high demand (Tableau & Forrester, 2022), and data visualization literacy is important for being an engaged citizen (Börner et al., 2019; Glazer, 2011).

As early as 1966, and more widely since the 1980s, it has been argued that data visualization literacy—sometimes called graphicacy—is as important as both textual literacy and math literacy (Balchin & Coleman, 1966; Börner et al., 2019; Friel et al., 2001; Glazer, 2011; M. J. Padilla et al., 1986). Indeed, math and science education standards in the US include objectives pertaining to data visualization literacy (Common Core State Standards Initiative, 2010; NGSS Lead States, 2013), and recent legislation introduced in the US House of Representatives is aimed at increasing funding for data science education more broadly (Data Science and Literacy Act of 2023, 2023).

However, interpreting data visualizations is not easy. In the present chapter I focus on graphs, a type of visualization that represents relationships between variables using visuospatial features and the coordinate system (A. R. Fox, 2023). In a recent review of the graph comprehension literature, Fox (2023) astutely pointed out a conceptual paradox at the heart of this research. Graphs seem to communicate information effortlessly, relying on our visual system to rapidly extract patterns and easily see important relationships, rather than having to read words or look at numbers (Ciccione et al., 2023; Larkin & Simon, 1987; Szafir et al., 2016). To the contrary, however, research from the past 40 years has shown that the process of graph comprehension is laborious and error-prone, and has documented the many difficulties that children and adults face when mastering these skills (Börner et al., 2016; Dimara et al., 2020; A. R. Fox, 2023; Friel et al., 2001; Glazer, 2011; Leinhardt et al., 1990; Shah & Carpenter, 1995; Shah & Hoeffner, 2002). To make matters worse, despite efforts to increase data literacy via formal education, the most recent National Assessment of Educational Progress in the US revealed declines in data literacy skill between 2019 and 2022, a trend that had started before the pandemic (Drozda, 2023; U.S. Department of Education, 2022).

One feature of graphs that contributes to this difficulty is that they are abstract simplifications of the world; to use them effectively, a cognizer must map the marks on the page—the lines, colors, and spatial positions—onto the real-world information that these visuospatial features are intended to represent (Hegarty, 2011; Tversky, 2011). This deep relational structure, both in what graphs aim to communicate and how that information is represented, can be difficult to extract, particularly for novices. Despite evidence that this mapping process is a stumbling block during graph comprehension, the domain-general cognitive capacity termed relational reasoning, which supports making abstract mappings, has yet to be fully explored in the context of graphs.

The present work consists of two complementary parts. In the first part, I propose that research on graph comprehension, including difficulties, cognitive models, and pedagogy, would be strengthened by considering relational reasoning as a foundational cognitive skill. To do this, I review literature from cognitive psychology and education and highlight various components of graph comprehension that are likely supported by relational reasoning. Further, I discuss how the lens of relational reasoning could be used to improve graph comprehension. In the second part, I begin to test these claims by conducting an empirical intervention study to investigate the potential benefits of explicitly engaging relational reasoning during graph learning. The study aimed to improve middle school students' understanding of graph concepts and problem-solving through a 2-day lesson and focused on graphs depicting linear functions, one of the simplest types of graphs since the relation between x and y is represented explicitly with a line. Before delving

into the first section, I begin by introducing relational reasoning and the related concept of relational complexity.

Relational reasoning and relational complexity

Relational reasoning is the capacity to map abstract, generalizable relations between two or more objects or representations, and identify meaningful patterns in information (Alexander, 2016; Holyoak, 2012). For example, a cognizer can map simple relations between surface-level features, such as noticing that two circles are similar because they are both red (first-order relations), as well as map higher-order relations between relations, such as mapping the relational structure of an atom's nucleus and its electrons to the relational structure of a sun and its planets (Gentner, 1983). This capacity has been shown to underlie abstraction, generalization, analogical reasoning, and fluid reasoning, and is considered core to human cognition (Gentner, 2003; Halford et al., 2010; Hofstadter, 2001; Starr et al., 2023).

Though relational reasoning is considered a powerful mechanism for learning (Cattell, 1987b; Gentner, 2003; Goswami, 2001; Hummel & Holyoak, 2005), it can also serve as a processing bottleneck. Specifically, cognitive load is influenced by relational complexity: the number of variables that need to be related and integrated to make sense of information (Andrews & Halford, 2002; Halford et al., 1998). As the number of variables increases, so too does the load, and therefore the processing time and instance of errors. It has been estimated that four variables is the maximum number that can be integrated in a single processing step without invoking other strategies, such as breaking down the step into smaller parts (Halford et al., 1998). Converging evidence from cognitive, education, and neuroscientific studies show that relational reasoning develops slowly over childhood and into early adolescence, reaching adult-like levels around 11 to 12 years old (Andrews & Halford, 2002; Crone et al., 2009; Jablansky et al., 2016; Wendelken et al., 2017).

Research on relational reasoning has stressed its importance for and applications to scientific reasoning and inquiry (Dumas, 2017; Klahr et al., 2013; Murphy et al., 2017; Resnick et al., 2017), education broadly (Dumas et al., 2013; Richland & Simms, 2015; Vendetti et al., 2015), and STEM education specifically (Alexander, 2017; Murphy et al., 2017). As discussed below, the ways in which relational reasoning capacity can be scaffolded or leveraged to support STEM education is a nascent, yet active, area of research, and it has yet to be fully explored in the context of graph comprehension.

4.3 A relational reasoning perspective on graph comprehension

I propose that applying the lens of relational reasoning to graph comprehension would strengthen research and pedagogy. The abstract nature of graphs means that a mapping process relating the visuospatial external representation to an internal representation is necessary to make meaning of the graph. Indeed, some graph comprehension researchers have gone as far as to say that these visuospatial properties are metaphors or analogies for the real-world properties and relations they represent (Hegarty, 2011; Shah et al., 2005). Just as words and concepts bind to roles in metaphor and analogy, real-world information must be bound to the visuospatial properties, patterns, and axes of the graph.

Moreover, these bindings are dynamic, in that they differ between graphs. For example, two graphs that show the exact same visual pattern can take on different meanings and be used to represent different relations depending on the axis labels and legend. Though some researchers have posited that abstract reasoning ability, another term for relational reasoning, plays a role in graph construction and interpretation, and that it may account for the difficulty that students have when using graphs (M. J. Padilla et al., 1986), few studies, if any, have explicitly tested this link. Below, I point to various ways in which relational reasoning may support graph comprehension, and then discuss how these insights could be applied to improve graph pedagogy.

Relational reasoning in the cognitive processing of graphs

Most task analyses and cognitive models of graph comprehension feature three main steps that have been referred to as (1) pattern recognition, (2) interpretation, and (3) integration (P. A. Carpenter & Shah, 1998; Freedman & Shah, 2002; Hegarty, 2011; L. M. Padilla et al., 2018; Pinker, 1990; see A. R. Fox, 2023 for a review). As a result, there are three main steps at which the design of the graph and the processing of the cognizer can cause bottlenecks for understanding. In the first step, the cognizer must identify and encode the visual patterns displayed in the graph. For example, an individual may notice that a set of points are clustered together or notice the steepness and direction of the trend of the points or of a line on the graph. Research suggests this first step happens quickly and with relative ease (Ciccione et al., 2023; Franconeri et al., 2021; Szafer et al., 2016). Steps two and three are where higher-order relational reasoning comes into play, when the cognizer begins making comparisons and mapping these visual features to meaning.

In the second step, the cognizer translates the visual patterns perceived in the graph into the quantitative and qualitative conceptual relations they represent. In relational reasoning terminology, the cognizer is mapping relations between the external representation and internal concepts. For example, she may map that the different colors of points represent different groups and that a line rising from left to right indicates a positive relation between the two variables of interest. Note that these quantitative or qualitative interpretations must be either already learned, and therefore retrieved from prior knowledge, or otherwise inferred from the graph (P. A. Carpenter & Shah, 1998), a process that relational reasoning would also support.

Finally, in the third step, the cognizer integrates these visual and conceptual patterns with their real-world referents by interpreting the patterns in the context of the information provided in the axis labels, legend, and title. In other words, she must map higher-order relations between the observed patterns of visuospatial relations and the patterns of relations between the real-world referents. This stage is critical because it is what makes a graph meaningful and useful for communicating information rather than simply looking at lines, colors, and patterns on a page. In an eyetracking study, Carpenter & Shah (1998) found that participants spent more time looking back and forth between the visual pattern and the axis labels and looking at the labels, legend, and title than looking at the visual pattern alone. They interpreted this gaze pattern, or visual routine, as suggesting that it is difficult to keep in mind information about the referents. However, this gaze pattern could have another, complementary interpretation. Past eyetracking studies of

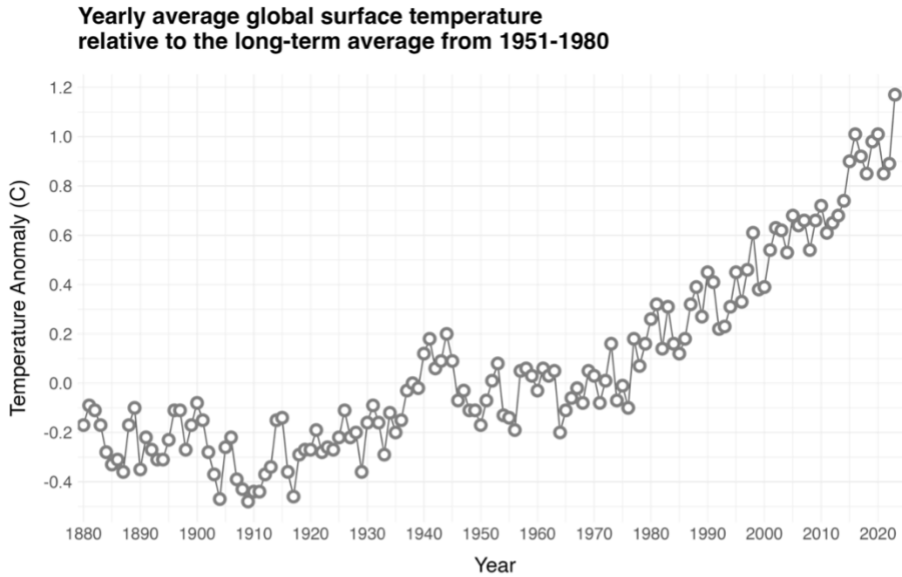
relational reasoning have interpreted saccades between visual stimuli as evidence of integrating the relations displayed in each of the stimuli—i.e., relational integration (e.g., Chen et al., 2016; Guerra-Carrillo & Bunge, 2018). Through this lens, results from the eyetracking study of graph comprehension (P. A. Carpenter & Shah, 1998) can be interpreted as suggesting that a bulk of processing time is spent integrating the visual features with their real-world referents, and that this step of graph comprehension is the most cognitively demanding. Dovetailing with this interpretation, Michal and Franconeri (2017) found that gaze patterns both reflect and affect the relations that are extracted from a graph and can lead to different graph interpretations.

Examining these three processing steps through the lens of relational reasoning reveals various places where this domain-general cognitive capacity could support graph comprehension. Despite the importance and cognitive difficulty of the last two steps, most of the research from psychology has focused on the pattern recognition and encoding step, investigating how visual features are perceived and how to design graphs more effectively to facilitate identifying the relevant visual patterns (for a review, see Fox, 2023 and Franconeri et al., 2021). Comparatively less research has focused on the difficult process of mapping the higher-order relations and how to improve performance on these later steps of graph comprehension, though some researchers have offered design suggestions aimed at minimizing these difficulties (e.g., Kosslyn, 1989; Matlen et al., 2020). Additional research is needed to better understand the relational structures that individuals use to represent and reason about the information contained in graphs, and to measure the relational complexity and cognitive load of these processing steps. The next section discusses the levels of questions that cognizers may be asked about graphs and their associated difficulties.

Relational reasoning and the levels of graph comprehension difficulty

Researchers in education and related fields approach graph comprehension from a different angle and have focused on the types of questions that can be asked about graphs and their associated difficulties. They have coalesced around three levels of graph comprehension questions, which increase in difficulty and build on each other (for a review, see Friel et al., 2001). These levels can be considered either the types of information that can be extracted from the graph or the level of understanding that a cognizer has about graphs. Here I propose that each of the three comprehension levels requires increasingly relationally complex computations, contributing to the increasing difficulty.

The first and easiest level of questions is termed “elementary”, or “reading the data”, asking the cognizer to extract a specific value from the graph (e.g., What was the temperature anomaly in 2000?; Figure 4.3.1). At this level, the cognizer is using the graph like a lookup table, and, notably, the information needed to answer this question can be found at one location (e.g., the height of a point on a line). Though answering this question requires all three steps of cognitive processing described above, the mapping problem is relatively simple because the value of the answer is displayed directly in the graph (Figure 4.3.1). Accordingly, students experience relatively few difficulties with questions at this level (Friel et al., 2001).



Levels of question difficulty	“Reading the data” Extract point-based information	“Reading between the data” Compare, integrate, and interpret to find relationships	“Reading beyond the data” Extend, predict, generalize, or make inferences
Example question	What was the temperature anomaly in 2000?	How did the temperature anomaly change between 1980 and 2000?	Describe the general trend in temperature anomalies. Based on this trend, what temperature anomaly do you predict for 2025?
Task analysis based on cognitive models and relational reasoning perspective	<p>Step 1: Visually locate the point for the year 2000 and move over to the y-axis to read the value.</p> <p>Steps 2 and 3: Map the y-value, .4, with the y-axis label, Temperature Anomaly</p>	<p>Step 1: Visually scan the points between 1980 and 2000.</p> <p>Step 2: Map that “going up” means “increasing”</p> <p>Step 3: Integrate these visual and conceptual patterns with the axis labels to infer that the global temperature increased over this range</p>	<p>Step 1: Visually scan all the points, noticing that over some ranges the points are moving up faster than in other ranges; also notice the “up and down” pattern of the points within closeby years</p> <p>Step 2: Map that ranges that “go up” faster are increasing faster, and compare this rate to other ranges; map the “up and down” pattern to indicating variability between years</p> <p>Step 3: Integrate these visual and conceptual patterns with the axis labels to infer that the rate of yearly global average temperature increase seems to have increased over time. Extend this pattern, taking variability into account, to predict that in 2025 the temperature anomaly will be ~1.1C</p>
Proposed relational demand and relational complexity	Low	Medium	High

Figure 4.3.1. Graph with examples of three levels of questions that can be asked about it (Friel et al., 2001) and a task analysis of the steps required to answer each question. This task analysis is based on cognitive models of graph comprehension (for a review, see Fox, 2023) and the relational reasoning perspective described here. Bolded words in the task analysis indicate processing steps that likely require relational reasoning. The proposed relational demand and relational complexity is also graded for each question level. The data plotted in the graph were retrieved from NASA.

The second level of questions—“intermediate” or “read between the data”—requires the cognizer to find relationships in the data, such as reasoning over a range of data (e.g., How does the temperature anomaly change between 1980 and 2000?; Figure 4.3.1) or comparing two points (e.g., Was the temperature anomaly greater in 1940 or 1950?). To answer these questions, the cognizer must integrate information from several locations on the graph (Bertin, 2001) and make at least one inference to get from the question to the answer (Curcio, 1987), resulting in more general statements about the graph. This level is more relationally complex than the first because information from more than one location on the graph must be integrated to generate an answer. Further, questions at this level have been shown to be more difficult than those at the first level (Friel et al., 2001). For example, students sometimes confuse an interval and a point, reporting information from a single point instead of integrating over a range of values (Leinhardt et al., 1990).

The third level of questions—“advanced” or “reading beyond the data”—requires the cognizer to extrapolate from the data and analyze the relationships presented in the graph to make generalizations, inferences, and predictions (e.g., Describe the general trend in temperature anomalies. Based on this trend, what temperature anomaly do you predict for 2025?; Figure 4.3.1). To answer these questions effectively, the cognizer must integrate over most or all of the data presented and understand the deep relational structure of the data (Wainer, 1992). Thus, these questions are the most relationally complex and require mapping higher-order relations between relations, as the cognizer makes a series of inferences and then integrates their output to generate an answer. Questions at this level are even more challenging and error-prone than at the intermediate level (Friel et al., 2001), and students often struggle with synthesis and coordination of evidence (Wilkerson & Laina, 2017). Considering these three levels together, I argue that the association between relational complexity and increasing question difficulty suggests that relational processing is a bottleneck for graph comprehension.

Relational complexity in graphical displays

Relational reasoning is also taxed by the complexity of the information being represented by the graph. One dimension of data complexity, sometimes referred to as graph complexity, is the number of variables being displayed on the graph. For example, in their eyetracking studies, Carpenter & Shah (1998) used 3-variable line graphs as stimuli and manipulated graph complexity in two ways, first by varying the number of lines on the graph (e.g., 2 vs 4 lines), and second by varying whether the lines had the same x - y relation (e.g., both positive slopes) or not. They argued that increasing complexity increased the number of inferences that needed to be made as well as the difficulty of the inferences. Note that the line on a line graph represents the relation between x and y , and the direction and steepness of the line’s slope represent the nature of that relationship. Slope is a binary relation because it relates two variables in one processing step. Specifically, it represents the change in y for every 1 unit change in x . When a second line with a significantly different slope from the first is added to the graph, showing an interaction, three variables must be integrated to interpret the graph: the relation between x and y now depends on a third variable, z (e.g., Figure 4.4.1, bottom panel). Therefore, increasing graph complexity increases the relational complexity.

Another source of complexity comes from representations of uncertainty, including variability, noise, and error (Franconeri et al., 2021). Whereas the line graphs in Carpenter & Shah (1998) depict linear functions, explicitly represent the x - y relation with a solid line, and do not show individual data points, a cognizer viewing the graph in Figure 4.3.1 would need to infer over all the data points to interpret the relation between x and y . Thus, graphs that depict uncertainty—such as by showing individual data points or by including confidence intervals or error bars—also require more inferences, increasing relational complexity.

Results from empirical studies provide evidence that increased graph complexity causes bottlenecks in graph processing. Carpenter and Shah (1998) found that as graph complexity increased across both of the dimensions that they manipulated, so too did the amount of time needed to process those graphs. Perhaps surprisingly, even seemingly simple graphs can require many comparisons. As Franconeri et al. (2021) point out, a simple bar graph with two bars each for two groups (e.g., 2×2 factorial design) for a total of four bars results in six possible pairwise comparisons that could be made in addition to two main effects and a possible interaction! Recent work has shown that the more comparisons that need to be made, the slower and more error-prone the processing (Franconeri et al., 2021; Nothelfer & Franconeri, 2020). Together, these studies suggest that although relational reasoning supports graph comprehension, relational complexity imposes processing constraints, resulting in higher cognitive load, longer processing times, and sometimes errors during graph interpretation.

Implications for graph pedagogy

Despite the critical role of mapping and inference at various points in the graph comprehension process, the direct link between relational reasoning and graph learning has yet to be explicitly tested. Further, most graph comprehension studies in psychology investigate adult processing, and relatively few have focused on children during learning. Those that have studied children have mostly focused on perceptual components (Ciccione et al., 2023; Kaminski & Sloutsky, 2013; Michal et al., 2018). In one relevant study that discusses relational structure and mapping, Gattis (2002) investigated whether non-spatial concepts are intuitively mapped onto space in systematic ways (e.g., cognitively constrained) or based on arbitrary conventions (like the arbitrary shapes of letters). She found that 6- and 7-year-old children in Germany with no graphing experience systematically mapped quantity to height and rate to slope, dovetailing with results from adults (Gattis & Holyoak, 1996). However, none of the graphs used in this study had axis labels, and it did not test how students were interpreting the visual patterns in the context of real-world referents. Another shortcoming of the graph learning literature is that most studies are cross-sectional, and do not aim to capture change over time or as a result of an intervention (Glazer, 2011).

I propose that relational reasoning can be scaffolded to support learning about graphs. Even though graphs are inherently relational, it is not clear that students and novices are aware of this structure or explicitly taught to understand graphs in this way. As diSessa et al. (1991) explain in the discussion of their study on 6th grade students learning to graph, “one of the difficulties with conventional [graph] instruction... is that students' meta-knowledge is often not engaged, and so they may come to know "how to graph" without understanding what graphs are for or why the

conventions make sense.” For example, graphing has traditionally been taught prescriptively, telling students to create certain graphs in some situations and other graphs in different situations without attention to how to decide which graph to create or even why one should use a graph in the first place (Friel et al., 2001), though this may be changing as the focus on data literacy becomes more prevalent in classrooms. Additionally, Glazer (2011) argues that graph comprehension cannot be learned “by osmosis,” but rather needs to be explicitly taught. I interpret this statement as having two meanings. First, merely being exposed to graphs does not mean that students understand how or why they are used; rather, their deep relational structure must be explicitly taught. Second, just because a student can see or identify a visual pattern does not mean that they understand what it means.

The lens of relational reasoning suggests three main ways to enhance graph comprehension. First, relational reasoning can be explicitly engaged to draw students’ attention to the relational nature of graphs, which could help elucidate why and how graphs are used. For example, students could compare instances of when it is useful (or not) to create a graph to better understand the contexts that would benefit from this tool. As another example, students could look at a series of different types of graphs and be prompted to simply state what the goal of each graph is—that is, what relations the graph aims to communicate. A second way that relational reasoning could be scaffolded is by having students explicitly practice mapping relations between the visual patterns and their conceptual and real-world referents. Third, prior research suggests that comparison is a powerful tool for engaging relational reasoning processes (see Vendetti et al., 2015 for a review), and this may extend to the context of graph comprehension. Typically, comparisons are encouraged between two visual representations presented side by side (e.g., two graphs, two diagrams, or two images). However, relational processes could also be elicited by asking cognizers to make comparisons between various features of the graph at various levels of abstraction, from comparing two points to comparing the structure of two lines to comparing the meaning of two relations. Although comparison questions are commonly used to test graph understanding, they have not to our knowledge been studied for the purposes of learning how to interpret and make meaning from graphs. Taken together, this section suggests that graph comprehension pedagogy would be strengthened by scaffolding relational reasoning.

4.4 Preliminary study

In this section, I begin empirically testing the direct link between relational reasoning and graph learning. To do so, I designed an intervention study that manipulated the extent to which two well-matched lessons on line graph concepts emphasized relational reasoning. In a pretest-posttest design, students were randomly assigned either to a lesson emphasizing encoding the visual features of the graph and point reading (VF lesson) or to a lesson emphasizing higher-order relational reasoning via mapping the visual features to their real-world referents and generalizing over ranges (RR lesson). The present study focuses on interpreting graphs that depict linear functions, which can be considered a building block for interpreting graphs visualizing real-world data. Whereas data graphs have complexities like variability and error that require more inferences to interpret, linear function graphs explicitly represent the relation between x and y with a line. That said, there are still many relations that must be encoded, integrated, and mapped in order to interpret linear function graphs, as will be described. Thus, designing a lesson on linear

function graphs is a straightforward starting point for investigating the link between relational reasoning and graph learning, and the present study serves as a proof-of-concept that graph lessons can be designed to scaffold relational reasoning.

I worked with US children in grades 5 to 7 (ages 10-13) to capture early graph learning and be able to impact understanding as it is developing. During these grades there is a shift in the math and science standards from working with more simple graphs that represent discrete variables only to working with graphs that represent continuous variables (Common Core State Standards Initiative, 2010; NGSS Lead States, 2013), a shift that has been shown to be difficult for students (Boote & Boote, 2017). I chose the concepts of y-intercept and slope because they are developmentally appropriate and generalizable across math, science, and statistics. Slope first appears in the Common Core Math 7th grade standard about ratios and proportional relationships, and y-intercept is first mentioned in an 8th grade standard about the linear formula, $y = mx + b$ that also mentions slope (Common Core State Standards Initiative, 2010). Further, students struggle with understanding the meaning of slope (Glazer, 2011). Critically, both concepts are relational and have visual definitions as well as contextual meanings. Y-intercept, the value of y when x is 0, relates x to y at a single point, whereas slope, the change in y for each unit of x, is a rate and describes the nature of the relation between x and y integrating over all the points, making slope a higher-order relation than y-intercept (Gattis, 2002).

In the VF lesson, participants' attention was drawn to individual points on the graph and were prompted to focus on the visual patterns (e.g., how steep the line looks) rather than being prompted to map these visual patterns to the context of the variables. On the other hand, in the RR lesson, participants' attention was drawn to ranges and trends on the graph and they were prompted to map the patterns that they saw to the real-world referents that they represented. For example, the VF lesson explains positive slope by saying, "A positive slope means that the line is going up from left to right," focusing students on the visual aspect of the slope, whereas the RR lesson explains, "A positive slope means that as x increases, the value of y also increases," focusing students' attention on relating x to y and the relational nature of the slope.

After the content part of the lesson, students saw four additional practice graphs with corresponding practice questions that maintained the focus of the assigned lesson condition. The practice stimuli featured graphs with two lines on them and practice questions encouraged students to make comparisons between the lines, a domain-general process that has been shown to engage relational reasoning during learning (see Vendetti et al., 2015 for a review).

The pretest and posttest had two parts. In the first part, students were tested on their understanding of the concepts of y-intercept and slope by being asked to explain them. The second part was a problem-solving task adapted from Moon et al. (2018). Students were presented with a graph and a series of questions that required integrating information from the graph with their reasoning to answer effectively. The graph for these questions featured three lines (i.e., the z variable had three levels), a level of complexity that the students had seen before, therefore increasing the novelty and difficulty of the task. We also assessed students' domain-general relational reasoning using a matrix reasoning task.

I predicted that matrix reasoning score would positively predict initial understanding of the concepts of y-intercept and slope as well as problem-solving performance with graphs. Next, I predicted that students who participated in the RR lesson would develop a deeper understanding of y-intercept and slope, as measured in the posttest, than students who participated in the VF lesson. Further, after the lesson, I predicted that students in the RR lesson would more effectively solve the problem-solving transfer task because the lesson scaffolding higher-order relational reasoning would better equip them to work with the complex graph, even though participants in neither condition had explicitly been taught how to solve this type of problem during the lesson. Finally, I tested whether matrix reasoning score predicted the magnitude of change in graph scores from pretest to posttest, and whether this effect depended on lesson condition. This test was exploratory in that I did not have predictions about the direction of the effects because multiple outcomes were plausible. For example, it is possible that students with lower relational reasoning may benefit more from a lesson that draws their attention to the visual features (VF lesson) or they may benefit more from a lesson that scaffolds their relational reasoning (RR lesson), which may otherwise be difficult to do without external support. In addition to testing these concrete predictions, this preliminary study also served as a proof-of-concept to determine whether and how the lens of relational reasoning could be applied concretely to design instructional materials for graph learning.

Methods

Participants

A total of 287 US students in grades 6, 7, and 8 ($n_5=39$, $n_6=42$, $n_7=206$) participated in May 2021. The study did not collect information on participants' age, gender, race, ethnicity, or socioeconomic status. The 6th grade participants were recruited from two class periods taught by the same math teacher at a public school in the San Francisco Bay Area. The 5th and 7th grade samples were recruited from the lower middle school (5th/6th) and upper middle school (7th/8th) of a public school district in the Greater Pittsburgh Area. The 5th grade participants were from two class periods taught by the same math teacher. The 7th grade participants were recruited from 10 class periods taught by two different teachers, four periods from one teacher and 6 periods from the other. There were two different levels of math for the 7th graders in our sample—"7th grade math" and pre-algebra—and each teacher taught half of their periods at one level and half at the other level.

The study was performed in accordance with the Internal Review Board at the University of California, Berkeley. Before the study, teachers sent a letter to students' parents/guardians from the researchers with information about the study and how to opt their child out if they did not want their child to participate. Students whose parents opted them out or who did not want to participate were given other classwork to complete during the class period. All participants verbally assented to participating.

The 6th grade sample, which was collected first, was used qualitatively as a discovery sample to develop the rubrics for scoring the open-ended questions. The 5th and 7th grade samples were the experimental samples. All reported analyses are with data from the 5th and 7th grade participants.

Procedure

The study took place over two consecutive days in students' math classes during a full class period. The 6th grade sample participated remotely over Zoom (50-minute periods) and the experimenter joined their Zoom classroom. The 5th and 7th grade students were in-person in their math classrooms, and the experimenter joined the class via Cisco WebEx to administer the study (5th grade: 88-minute periods; 7th grade: 40-minute periods). All tasks were administered on Qualtrics and students participated on their own computers with headphones. Each day of the study, students were emailed a unique URL to access the study materials at the beginning of their class period. While working through the tasks, once students clicked the "next" arrow to advance to the next page, they were not able to return to the previous page or question.

There were five parts of the study: basic graph knowledge assessment, pretest, graph lesson, posttest, and a matrix reasoning task. Participants were randomly assigned to one of two well-matched lessons on y-intercept on slope: (1) a relational reasoning lesson or (2) a visual features lesson. Random assignment was done within classrooms. The basic graph knowledge assessment, pretest and the first half of the lesson (instructional block + first two graphs of practice block) were administered on the first day, and the second half of the lesson (last two graphs of practice block), the posttest, and the matrix reasoning task were administered on the second day. Participants were given until the end of their class period to complete the tasks assigned for each day. At the end of the class period, participants were asked to exit out of the Qualtrics survey, regardless of whether they had finished. What part students started with on the second day depended on where they left off on the first day. If after the first day students had not finished the first part of the lesson, they started the second day with where they had left off the prior day before continuing to the second part of the lesson. Otherwise, students started with the second half of the lesson at the beginning of the second day.

Tasks

Basic graph knowledge assessment. In this task, students were first shown a line graph and then were asked a series of six multiple-choice questions about the graph that were designed to assess whether students had the pre-requisite graph reading knowledge and skills to engage with the pretest and lesson (see the Supplementary Materials for the graph and full set of questions). The line graph's x- and y-axes were labeled "Hours studied" and "Test score (out of 100)", respectively, and the one line on the graph was labeled with a name ("Brianna"). The first two questions asked participants to identify the variable on the x- and y-axis, respectively, and gave each of the three variables as options as well as an "I don't know" option. The next four questions asked students to read points on the graph, and students responded by selecting their answer from a dropdown menu of numbers, which included all the numbers on both axes and "I don't know". Two questions gave the x-value and asked for the y-value, and two gave the y-value and asked for the x-value. Within each question type, one question was asked in the context of the axes (e.g., "If Brianna wants to score a 100 on the test, how many hours does she need to study?") and one was asked in terms of x and y (e.g., "What is the value of x when y = 85?"), yielding two with-context questions and two context-free questions.

Qualtrics automatically scored these responses, and if students answered five or six out of six correctly, they continued to the pretest. If students answered fewer than five of the six questions correctly (i.e., scored a four or lower), they were directed to extra graph reading practice. This extra practice included a video reviewing the axis labels and how to read points on a line graph. Then participants answered four additional point-reading questions in the same format as for the first graph with a new graph. This new graph had different axis labels, line, and line label (see Supplement). After completing the extra practice, students continued to the pretest, regardless of score on these extra questions. However, as I pre-registered, only students who answered all four of the extra questions correctly were included in analysis.

Pretest and posttest. The pretest aimed to capture what students already knew about the target concepts before completing the lesson and the posttest aimed to measure how their understanding and reasoning changed. The first part of the pretest and posttest measured knowledge of y -intercept and slope. The second part measured how students apply these concepts and higher-order relational reasoning in a problem-solving scenario. The posttest also included an exploratory third part that examined students' transfer of graph concepts to a science context, which is not analyzed in the present study.

For the study with 6th graders, the questions were not timed, and responses were required for all questions to advance to the next page. If a student did not want to answer a question, they could enter any characters into the text box (e.g., spaces) or choose to write "I don't know." Because many 6th graders spent too long on these questions at pretest and ran out of time for the lesson, the open-ended questions auto-advanced for the 5th and 7th graders to help them pace themselves and make sure they finished the pretest with enough time to complete the lesson. The length of time for each question was calculated from the 6th grade sample to make sure that participants were not rushed but also did not spend too much time. When the page auto-advanced on an open-ended question, whatever response was written in the text box was what was submitted as the response. Further, responses were no longer required for students to advance to the next page. If a question was still blank when a student clicked the "next" arrow, a window popped up notifying the student, "There is 1 unanswered question on this page. Would you like to continue?" with two options "Continue without answering" or "Answer the question." Students were encouraged to answer all the questions.

Part 1: y -intercept and slope. In part 1 of the pretest, students were presented with the same graph from the basic graph knowledge assessment and asked a series of three questions. Students started by seeing a video introducing the task. Then, all three questions were presented on the next page. For the 5th and 7th graders, the program auto-advanced from this page after 5 minutes. In the first question ("click-on-graph"), participants were asked to click on the graph where they thought the y -intercept was. The Heat Map question type was used in Qualtrics, and a small region of interest around the y -intercept was pre-specified on the graph as part of the question setup. If students clicked within that region, received a score of 1. If students clicked outside of the region, they received a score of 0. If students did not respond to the question, it was marked as no response. The second two questions were open-ended and asked students to explain what the y -intercept and slope on the graph "tells you" (see Supplementary Materials for

the full question text). For the slope question, they were also asked to describe the slope of the line. The y-intercept question had a max score of 3 and the slope question had a max score of 4.

The outcome measure for this first part of the pre/posttest was the sum of the points for the “click-on-graph” question (max: 1 point) and the two open-ended questions (max: 3 and 4 points, respectively). As described in the pre-registration and scoring guide, children did not always understand the expected response type on the novel “click-on-graph” question, and sometimes left it blank. However, a no response did not always mean they did not know the answer to the question. Instead of marking a no-response as incorrect, the score was calculated differently for these participants: the sum score was first divided by 7, the max score excluding the “click-on-graph” question, and then multiplied by 8 (max score with this question), to put the score on the same 0-8 scale as the participants who did answer that question.

Part 2: Problem solving transfer task. In part 2 of the pre/posttest, students watched a brief video that shared a cover story and a novel graph with three lines on it, and then asked a series of open-ended and multiple-choice questions that involved reasoning with information from the graph. This task and the pretest cover story and graph were adapted from Moon et al. (2018). At pretest, the student was told that a middle school’s student council was buying t-shirts for the 8th graders and that the student’s goal was to help student council decide which of the three t-shirt companies would be the cheapest option. They were shown a graph to help them make their decision, which had “Number of Shirts” on the x-axis, “Total Cost” on the y-axis, and three lines with varying slopes labeled as Company A, Company B, and Company C in the legend. The posttest cover story was analogous but with a different context: the student was told they were hosting a birthday party for their best friend and needed to decide what food to serve. They were shown a graph with “Party Attendance” on the x-axis, “Total Cost” on the y-axis, and the foods “Burgers,” “Pizza,” and “Spaghetti” labeled in the legend.

Students were then asked a series of four open-ended and two multiple-choice questions based on the graph. Critically, these questions asked students to reason about and integrate information presented in the graph in order to make a decision and give a recommendation about a course of action. The goal of these questions was to measure how sophisticated their reasoning was, which included capturing what kind of information students were attending to in the graph (e.g., point-based versus ranges versus slopes) and how they integrated this graph evidence into their reasoning. The questions were designed to start off more open-ended and open to interpretation to capture what students spontaneously generated (e.g., open-ended question 1: “Student council does not know yet how many students will buy a t-shirt. Make a recommendation to student council about which company they should use to make the t-shirts.”), and then get narrower and more specific to examine whether, when prompted, students would engage in more sophisticated reasoning, such as attending to ranges instead points if they did not already do so in the earlier open-ended questions (e.g., open-ended question 4: “When, if ever, would buying from Company B be the cheapest option?”). The two multiple-choice questions asked students to compare the companies and select the cheapest based on their y-intercept and slope, respectively, in the context of the cover story (y-intercept: “Which company has the lowest

starting cost?; slope: “Which company has the lowest cost for each additional t-shirt purchased?”).

Questions were presented one at a time, with one question per page. Each page showed the graph and the question below it, and, as in the rest of the study, students could not return to a question once they advanced to the next. For 5th and 7th graders, the open-ended questions auto-advanced as described above. The two multiple-choice questions were not timed, and responses were required to advance to the next page.

The outcome measure for this second part of the pre/posttest was the sum of points from the four open-ended questions and two multiple-choice questions. The open-ended questions were each worth a different number (5, 4, 3, 3, respectively) based on their difficulty, and the multiple-choice questions were each worth 1.5 points. Thus, the maximum possible score on part 2 was 18 points. See the scoring rubric for additional scoring details.

Open-ended question scoring rubrics. The open-ended question scoring rubrics aim to capture the sophistication of students’ reasoning in their responses. Sophisticated responses demonstrate conceptual understanding, complexity in graph interpretation skill, and higher-order relational reasoning, such as making higher-order comparisons, mapping between the visual properties of the graphs and the context of the axes, relating the x variable to the y variable, and considering and comparing ranges of values instead of individual points. The rubrics were developed through an iterative process. First, version of the scoring rubric was based on the literature (e.g., Boote & Boote, 2017; Friel et al., 2001; Moon et al., 2018), learning goals, and pilot participant responses, and posted to OSF. Then, this rubric was used to score the 6th graders open-ended responses, who were the discovery sample. Since the 6th grade data were being used for the open-ended responses and not for analysis, no data cleaning procedures were applied and all given responses scored, even if students had not finished all the tasks. This scoring was completed by two independent coders. After this stage, the rubric was revised to account for the variation in responses and edge cases that were not captured in the original version. The scoring guide also outlined how the questions would be weighted (i.e., the max score for each question). After this updated version was posted to OSF, two independent coders scored the open-ended responses from the 5th and 7th grade students, and overall agreement was high ($r = .94$; see Supplementary Materials for agreement by question). Scoring disagreements were decided by the first author. All scoring was completed blind to lesson condition, and scorers alternated whether they scored the pretest or posttest first.

Lessons on y-intercept and slope. The two lessons on y-intercept and slope both covered the same content and had the same basic structure. Before describing the relational reasoning manipulation, I first describe the components that are the same between the lessons. The lessons were divided into two parts, the instructional block, which taught the concepts of y-intercept and slope, and the practice block, which consisted of a series of four graphs with associated practice questions. To begin the instructional block, students first saw a short video (28s) with general instructions about what they would be doing in that section. Next, participants were introduced to the main concepts of the lesson: 1) y-intercept, 2) slope direction, and 3) slope steepness. For

each concept, the participant first watched a brief video explaining the concept (min = 15s, max = 83s). How the concept was explained was manipulated by lesson condition, as described below. Then, participants answered multiple-choice questions to practice the concept and received feedback before moving onto the next concept. The wording of these practice questions and the feedback given also depended on the lesson condition, but the graphs and answers were the same in both lessons. Given how brief the instructional block was—the longest video was only 1 minute and 23 seconds—the same two contexts were used for all the videos and questions, either the relation between hours studied and test score or between hours of tv and hours of sleep. Some of the graphs showed one line and others showed two for two different people. In the instructional block, the graphs with two lines always had either the same slope but different y-intercepts or the same y-intercept but different slopes, to highlight the concept that was being featured in that part of the lesson.

After completing the instructional block, participants began the practice block. They saw a series of four graphs, each with five to six associated questions. The graphs were the same between lesson condition, but the wording of the questions depended on condition. Unlike in the instructional block where all questions had the same correct answers between conditions regardless of how they were worded, in the practice block the questions were well matched, but sometimes the correct answers were different between conditions. These practice graphs all showed two lines, and the lines now always had different y-intercepts and slopes. Three visually different interactions were represented among the graphs: crossing positive lines, crossing positive and negative lines, and non-crossing positive and negative lines. Since showing two lines with different y-intercepts and slopes was not introduced in the instructional block, the first graph of the practice block used the same hours studied-test score context as in the instructional block to scaffold the students to these more difficult graphs and questions. After the first graph, the other three contexts were all novel. Additionally, the first and second practice graphs showed a similar looking interaction—crossing positive lines—to help scaffold students to working with this type of graph in a novel context. The next two graphs showed different types of interactions and different contexts.

Relational reasoning engagement was manipulated between lesson condition by the way that concepts were explained and how practice questions and feedback were worded. The visual features (VF) lesson was designed to focus students' attention on the visual patterns present in each graph. On the other hand, the relational reasoning (RR) lesson was designed to help students practice mapping the visual patterns to their real-world referents. In the VF lesson, the concepts of slope and intercept were introduced by how they could be visually identified, whereas in the RR lesson they were identified by their meaning in the context of the x- and y-axis variables. For example, the VF lesson described the steepness of the slope as one line going up faster than the other line (Figure 4.4.1), drawing the students' attention to how the line looks on the page, but not to what that pattern means in the context of the axis variables. On the other hand, the RR lesson described steepness in terms of the relative change in the y variable for each one unit increase in the x variable for one level of the z variable compared to the other (Figure 4.4.1), drawing students' attention to the relation between x, y, and z and mapping the visual pattern to that relation. The graphs in the lesson videos were also animated differently to help focus

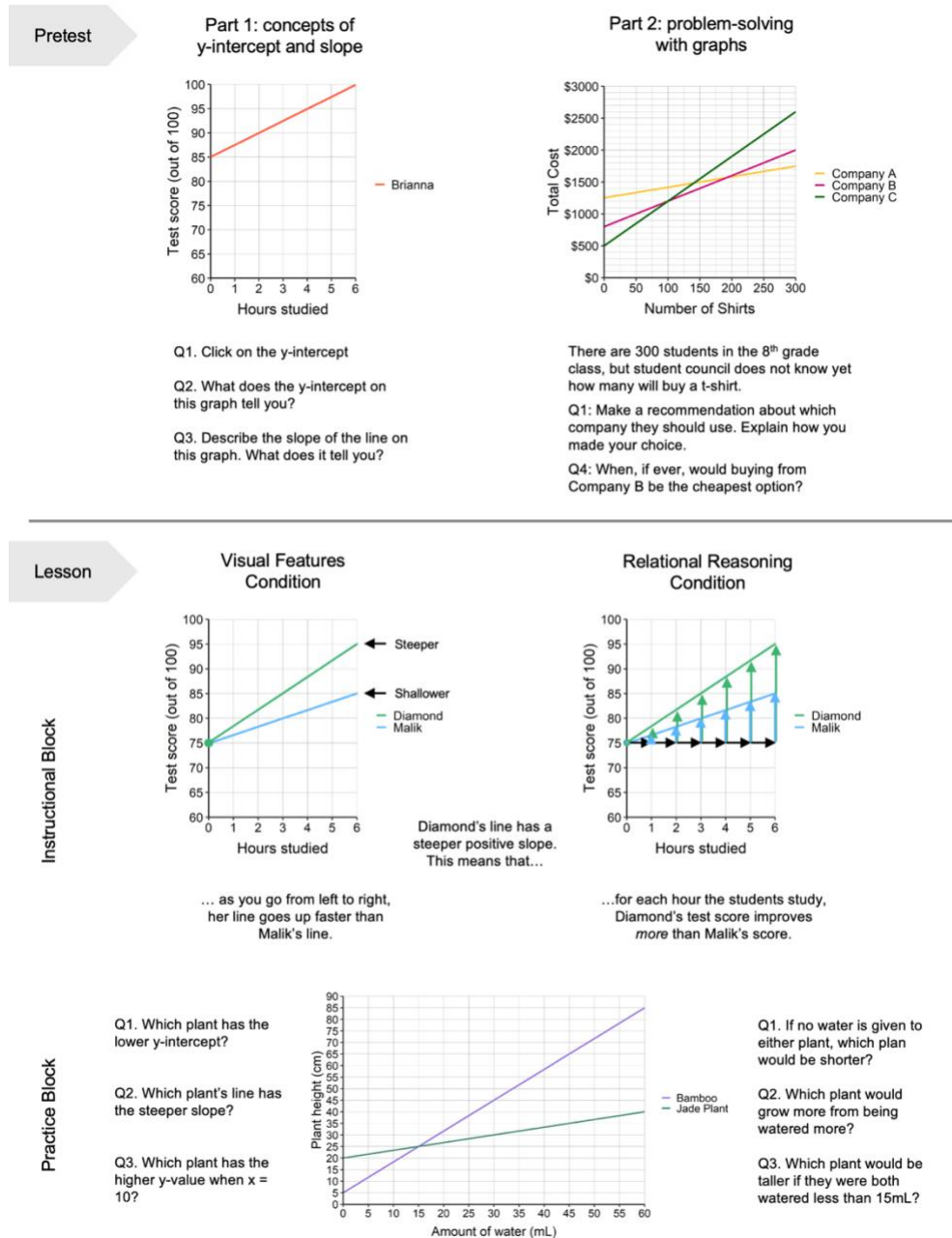


Figure 4.4.1. Pretest and lesson design with sample questions. The top panel shows the graphs used in the two parts of the pretest. Pretest part 1 measured students' understanding of y-intercept and slope with the three questions listed. Pretest part 2 measured students' problem solving with graphs by asking four open-ended questions, two of which are listed here, and two multiple-choice questions; this graph and cover story were adapted from Moon et al. (2018). The bottom panel shows sample instructional materials and practice questions from the two well-matched lesson conditions: Visual Features (left) and Relational Reasoning (right). The instructional block example shows how slope steepness was described differently in each lesson. The lines on the graphs were also animated differently between lessons. The lines on the VF graph slowly started appearing from left to right. On the RR graph, a black arrow appeared, then its corresponding green and blue arrows appeared, and the cycle repeated, to show the difference in the change in y for each one unit increase in x . The practice block shows the second practice graph with the first three questions that were asked about it in each condition.

students' attention to way the lesson was describing each concept (see Figure 4.4.1 for a snapshot of an instructional graph after animation, and videos on OSF to view animations in action).

In addition to differences in how the concepts were taught, the lessons also manipulated how students practiced them by using differently worded questions and giving feedback on their responses that reinforced how the concepts had been taught in that lesson. The VF condition had students practice reading points on the graph, and sometimes comparing point values when there were two lines on the graph. For example, for a graph in the practice block with two lines (Figure 4.4.1), students in the VF lesson were asked which line had the lower y-intercept and which plant had the higher y-value when x equaled 10. On the other hand, the RR lesson had students practice interpreting the concepts in the context of the graph. For example, for that same graph in the practice block, students in the RR lesson were asked "which plant would be shorter" if no water were given to either plant, making the students integrate the meaning of the y-intercept with the context of the graph (Figure 4.4.1). The RR lesson also engaged students in range-based reasoning by comparing lines over ranges to focus their attention on the overall structure of those lines, not just points. For example, students were asked to compare the two lines over the range of when x is less than 15 (Figure 4.4.1). Prior work has shown that students often struggle to interpret ranges on graphs (e.g., when asked "when is jade taller than bamboo?") and will often respond with a single point instead of a range (Moon et al., 2018; Swan & Phillips, 1998), likely because reasoning about a range is more relationally demanding since it requires integrating over more than one location. Thus, the RR lesson scaffolded this more complex graph skill by directing students' attention to ranges.

Finally, the RR lesson also drew students' attention to the structure of the lines through more general comparison questions. The first question for each new graph in the practice block of the RR lesson asked students to list the similarities and differences in the relation between the x and y variables for each level of the categorical variable (e.g., similarities and differences the relationship between amount of water a plant height for bamboo and jade). The goal was for students to practice seeing patterns and mapping them to meaning on their own before being prompted with questions for that graph. After answering the question, students saw a feedback page that listed some of the things they could have said to think about the graph in context, such as which plant has a greater starting point for height if the plants aren't watered all. The corresponding question in the VF lesson simply asked students to "explain to a friend what you see in this graph." In this question, students were not drawn to make any explicit comparisons and could give any level of explanation they wanted. The feedback said they could describe the y-intercept, the direction of the lines, and the steepness of the lines.

The responses from the questions in the instructional and practice blocks of the lessons were not scored or analyzed in the present study. See the Supplementary Materials for all graphs and questions; the instructional videos are posted on OSF. Though relational reasoning is being taxed in both conditions, higher-order relational reasoning should be more heavily taxed in the RR condition due to the focus on mapping visuospatial features to conceptual meaning.

Matrix reasoning task. This task was adapted from the Wechsler Intelligence Scale for Children IV matrix reasoning task (WISC-IV; Wechsler, 2003). The WISC-IV matrix reasoning task has 3 practice trials and 35 items. In the administration protocol, all participants start with all three practice trials, then children ages 9-11 start with item 7 and children ages 12-16 start with item 11. Since the study spanned these age ranges, I started all children on item 7 in order to compare scores between participants. Therefore, there was a total of 29 possible items (items 7 through 35). The items were scanned from the WISC-IV administration booklet and touched up for clarity using Photoshop.

The task was administered on Qualtrics. Participants saw a visual puzzle with one piece missing and five numbered images. They were instructed to “identify the missing piece” by selecting the corresponding image number. Participants were given feedback on the three practice trials. During the main task, participants were no longer given feedback and an additional answer choice, “I don’t know”, was added. Participants had a maximum time allotment of 90 seconds per item. If that maximum time was reached, the program marked the item as incorrect and auto-advanced to the next item. Items for which the participant selected the wrong choice, selected “I don’t know,” intentionally left blank, or ran out of time were marked as incorrect. If a participant responded incorrectly to four consecutive items (including timed-out items), then the task automatically ended, as per the WISC-IV matrix reasoning stoppage criteria. The outcome measure for this task was the total number of correct responses.

Data analysis

Data cleaning. A total of 39 5th graders and 206 7th graders participated in the study. One 5th grade participant, who participated remotely over Cisco WebEx, was removed because a parent was aiding him in answering the questions. Data were cleaned according to the pre-registration posted on AsPredicted (<https://aspredicted.org/blind.php?x=8xw5m7>). First, two 7th graders who did not start the lesson were removed from the data set (i.e., either did not finish the basic graph knowledge questions or did not finish the pretest). Next, participants who did not successfully meet the basic graph knowledge pre-requisite were removed. Participants who got more than 1 of the 6 questions incorrect on the basic graph knowledge questions at the beginning of the study watched a video reviewing how to read graphs³, then answered an additional four questions practicing reading points on a graph. Participants who did not answer all 4 additional graph practice questions correctly were removed from analysis ($n_5 = 9$; $n_7 = 31$). Participants were also excluded from analyses if they either did not complete the lesson ($n_5 = 1$; $n_7 = 23$) OR they did not engage with the lesson. Because students participated at their own speed on their personal laptop, it was difficult to monitor student engagement. Therefore, the time a student spent on the lesson was used as a proxy for engagement. I defined not engaging with the lesson as the conjunction of (1) spending less than 2.5 median absolute deviations (MADs) from the median time on the lesson AND either (2a) chance performance (50%) or worse on the multiple-choice questions or (2b) spending less time on the open-ended lesson questions than 2.5 MADs

³ The preregistration incorrectly stated that participants could get up to two incorrect on the basic graph knowledge assessment without seeing the review video. However, the task was programmed so that any participant who scored less than 5 correct was shown the video and given the extra assessment questions.

Table 4.4.1. Sample size by grade and analysis

Analysis	Grade 5	Grade 7
All analyses with matrix reasoning as a predictor	23	92
part 1 scores ~ timepoint (pre or post) * lesson condition (VF or RR) + 1 participant	28	147
part 2 scores ~ timepoint (pre or post) * lesson condition (VF or RR) + 1 participant	28	134

from the median time. All medians and MADs were calculated within grade. No participants were removed for not engaging with the lesson. Next, three 7th grade participants were removed for not starting the posttest. Additionally, participants would have been excluded if their time on any of the pretest or posttest sections was less than 2.5 MADs from the median time on that section, suggesting that they had rushed through the materials and had not engaged fully with them, but no participants met these criteria.

To retain as many participants as possible for the analyses, participants needed to have completed just the first part of the pretest to be included in any analysis and were included in the analyses that they had full data for. For example, if a participant completed part 2 of the posttest but not the matrix reasoning task, then that participant would be included in the analyses of change in part 1 scores and change in part 2 scores, but not in the analyses involving matrix reasoning as a predictor. This was important because due to the limited class time, 13 7th graders were not able to finish part 2 of the posttest (5 in VF condition), and an additional 42 7th graders did not finish the matrix reasoning task (21 in VF condition). There were also five 5th graders who did not finish the matrix reasoning task (3 in VF condition). The final sample sizes for each analysis are shown in Table 4.4.1.

Analysis methods. Due to the large difference in sample size between 5th and 7th grade, I analyzed the data for each grade separately. I also analyzed the two parts of the pretest separately because they address different research questions. To test my first research question about whether there was a relation between matrix reasoning and performance on each part of the pretest, I fitted linear regression models predicting pretest scores from matrix reasoning score. Next, linear mixed effect models were fit to test the effect of lesson on graph score. Part 1 scores were predicted by fixed effects of time point (pre or post), lesson condition (VF or RR), and the interaction between them as well as a random intercept for participant. A model with the same fixed and random effects structure was also used to predict part 2 scores. Finally, to test whether matrix reasoning was related to the magnitude of change in scores from pretest to posttest, and possibly whether that relation depended on lesson condition, I fit a linear model predicting change in the score of each part of the pre/posttest from matrix reasoning, lesson condition, and their interaction.

Transparency and openness. The hypotheses and data cleaning and analysis plan for this study were pre-registered on AsPredicted (<https://aspredicted.org/blind.php?x=8xw5m7>). We report all data exclusions and manipulations. All graph stimuli and questions are presented in the Supplementary Materials and the instructional videos are on OSF. All data cleaning and analysis scripts are available on OSF. Analyses were conducted in R Version 4.2.1 (R Core Team, 2022), tidyverse Version 2.0.0 (Wickham, 2023) packages, including dplyr Version 1.1.3 (Wickham et al., 2023) and ggplot2 Version 3.4.4 (Wickham, 2016), were used for data wrangling and visualization, respectively. In addition to ggplot2, see Version 0.9.1 (Lüdecke et al., 2021), and gridExtra Version 2.3 (Auguie, 2017) were used to create the visualizations. lme4 Version 1.1.33 (Bates et al., 2015) and lmerTest Version 3.1.3 (Kuznetsova et al., 2017) were used to fit the mixed-effects models. papaja Version 0.1.2 (Aust & Barth, 2022) and english Version 1.2.6 (J. Fox et al., 2021) were used to format and output the results from R.

Results

Graph scores and matrix reasoning scores varied within grade

There was a great deal of variability in students' performance on the three assessment measures, as shown in Table 4.4.2. T-tests within grade confirmed that there were no differences in these measures between lesson conditions at pretest ($|t|s \leq 1.49, ps > .154$). Due to the difference in sample size, performance between grades is not directly compared, but visual inspection reveals that there are not large differences between grades. Additionally, even though according to the US Math standards 7th graders have had more experience with graphs than 5th graders, there were not ceiling effects for either grade, suggesting that on average 7th graders were still graph novices and could benefit from the graph lesson.

Matrix reasoning predicted higher graph scores at pretest

First, I tested the relation between matrix reasoning and initial understanding of the concepts of y-intercept and slope by fitting linear regression models for grades 5 and 7 (Figure 4.4.2, left panel). The model for grade 5 revealed that each additional correctly answered matrix reasoning question predicted a 0.20 point increase in the score on part one of the pretest ($b = 0.20, 95\% \text{ CI } [0.05, 0.35], t(21) = 2.81, p = .011$) and matrix reasoning score explained 27.28% of the variance in graph score on part 1 ($R^2 = .27, F(1, 21) = 7.88, p = .011$). However, for grade 7

Table 4.4.2. Descriptive statistics for each part of the pre/posttest and matrix reasoning.

Measure	Part 1: y-intercept and slope			Part 2: problem solving task			Matrix Reasoning
	Pretest	Posttest	Change	Pretest	Posttest	Change	
Grade 5	2.84 [0, 8] SD = 1.75	4.33 [1, 8] SD = 2.26	1.49 [-1.7, 7] SD = 2.09	6.16 [0, 15.5] SD = 4.33	9.05 [2, 17] SD = 5.05	2.89 [-1, 10.5] SD = 3.20	18.9 [7, 25] SD = 3.69
Grade 7	2.31 [0, 8] SD = 1.94	4.18 [0, 8] SD = 2.12	1.87 [-4, 8] SD = 2.37	6.85 [0, 18] SD = 3.95	8.74 [0, 18] SD = 4.64	1.86 [-10.5, 10] SD = 3.61	16.1 [0, 24] SD = 4.42

Note: Mean, range, and standard deviation; the maximum possible score for each measure is 8, 18, and 29, respectively; Change is the difference in score from pretest to posttest and was first calculated within individual then summarized

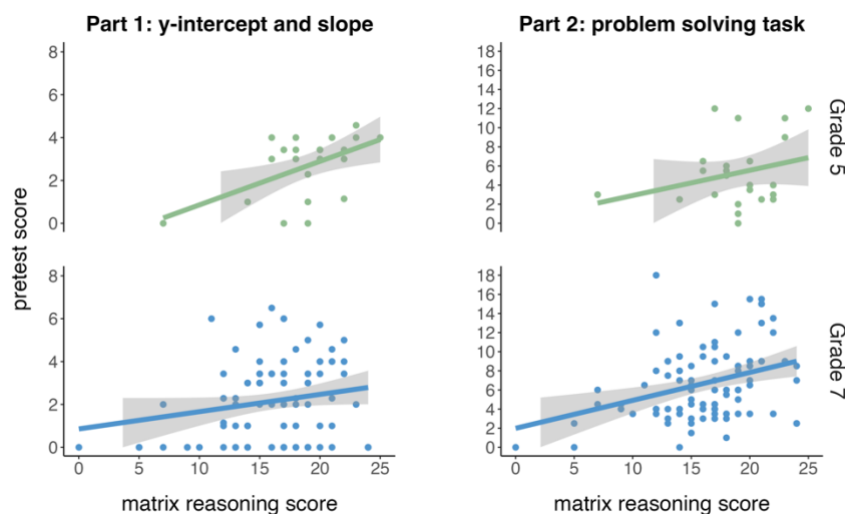


Figure 4.4.2. Relation between matrix reasoning score and graph score by grade for both parts of the pretest. In the left panel, the outcome measure is score on part 1 of the pretest, which assessed students’ initial understanding of the concepts of y-intercept and slope. In the right panel, the outcome measure is part 2 of the pretest, which assessed students’ problem-solving using graphs as evidence in reasoning. The gray portion shows the 95% confidence interval around the linear model.

matrix reasoning did not predict performance on part 1 of the pretest ($b = 0.08$, 95% CI $[-0.01, 0.17]$, $t(90) = 1.84$, $p = .069$).

Next, I tested the relation between matrix reasoning and initial performance on the problem-solving task by fitting linear regression models for grades 5 and 7 (Figure 4.4.2, right panel). The model for grade 5 did not reveal a significant relation between matrix reasoning score and performance on part 2 of the pretest ($b = 0.26$, 95% CI $[-0.15, 0.68]$, $t(21) = 1.31$, $p = .204$). However, matrix reasoning was a significant predictor of 7th graders performance on the problem-solving task at pretest: each additional correctly answered matrix question predicted a 0.29 point increase in the part two score ($b = 0.29$, 95% CI $[0.12, 0.47]$, $t(90) = 3.29$, $p = .001$) and matrix reasoning explained 10.73% of the variance in part 2 score ($R^2 = .11$, $F(1,90) = 10.82$, $p = .001$). It is worth noting that despite not finding a significant effect in grade 5, the effect sizes are comparable between the two grades (Grade 5: $r=0.27$; Grade 7: $r=0.33$), with overlapping 95% confidence intervals (Grade 5: 95% CI $[-0.15, 0.62]$; Grade 7: 95% CI $[0.13, 0.50]$), perhaps suggesting that our sample is underpowered to detect an effect of matrix reasoning in grade 5.

Lesson improved graph scores for both conditions

To test the research question about whether participating in a graph lesson that emphasized relational reasoning would affect students’ understanding of concepts of y-intercept and slope, I fit a linear mixed effects model predicting score on part 1 score from fixed effects of timepoint, condition, and their interaction, and a random intercept for participant. The grade 5 model revealed a significant effect of timepoint ($\hat{\beta} = 2.15$, 95% CI $[0.99, 3.31]$, $t(26) = 3.63$, $p = .001$), indicating that the scores generally improved from pretest to posttest, and this improvement did not depend on condition ($\hat{\beta} = -1.15$, 95% CI $[-2.68, 0.39]$, $t(26) = -1.47$, $p = .155$). On

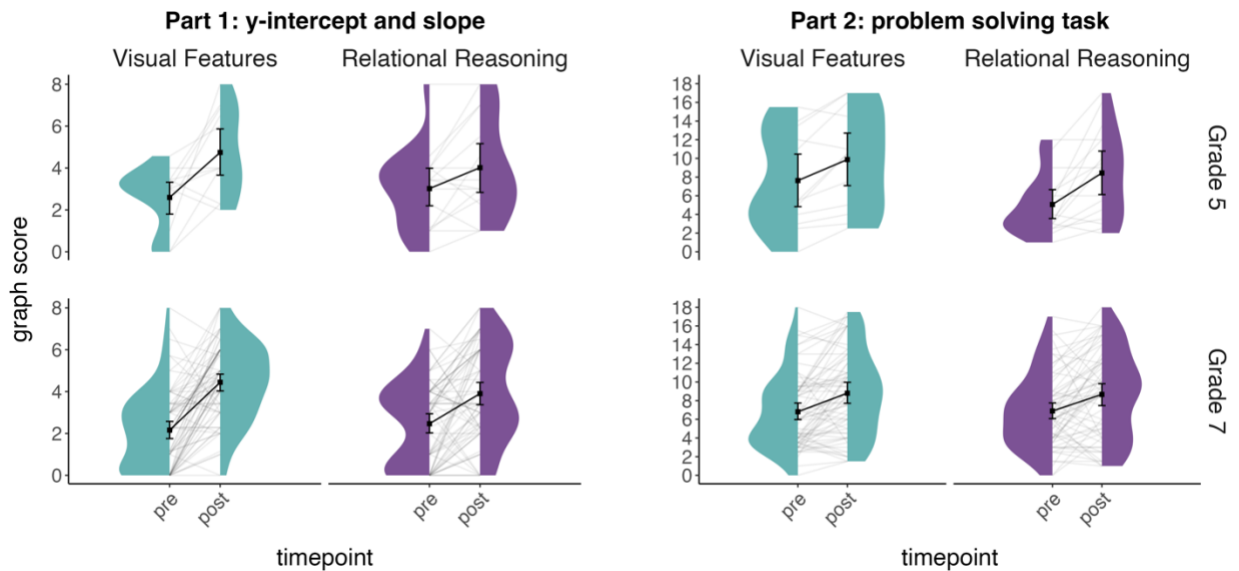


Figure 4.4.3. Change in graph score from pretest to posttest by lesson condition and grade. The left panel shows the results for part 1 of the pre/posttest, which assessed students’ understanding of the concepts of y-intercept and slope. The right panel shows the results from part 2 of the pre/posttest, which assessed students’ problem-solving using graphs as evidence in reasoning. Each light gray line represents a student, connecting their pretest score to their posttest score. The black squares represent the group means, and the error bars are bootstrapped 95% confidence intervals. The black line connects the group averages.

average, 5th graders improved 2.15 points out of 8 on the part 1 score from pretest to posttest after participating in either lesson (Figure 4.4.3).

The grade 7 model also revealed that the scores for students across both conditions improved on average from pretest to posttest ($\hat{\beta} = 2.28$, 95% CI [1.76,2.81], $t(145) = 8.52$, $p < .001$). However, for 7th grade, participants in the VF condition improved more than participants in the RR condition ($\hat{\beta} = -0.85$, 95% CI [-1.60, -0.09], $t(145) = -2.19$, $p = .030$). On average, 7th graders in the in the VF condition improved their scores by 2.28 points out of 8, whereas 7th graders in the RR condition improved by an average of 1.44 points out of 8 (Figure 4.4.3).

For the next research question about whether participating in the lesson that emphasized relational reasoning would improve students’ problem solving with graphs I again fitted a linear mixed effects model, now predicting part 2 score from the same fixed and random effects. The models for grades 5 and 7 both revealed that the scores for students across both conditions generally improved from pretest to posttest (Grade 5: $\hat{\beta} = 2.25$, 95% CI [0.44,4.06], $t(26) = 2.43$, $p = .022$; Grade 7: $\hat{\beta} = 1.94$, 95% CI [1.09,2.78], $t(132) = 4.50$, $p < .001$) and that this improvement did not depend on condition (Grade 5: $\hat{\beta} = 1.13$, 95% CI [-1.28,3.53], $t(26) = 0.92$, $p = .367$; Grade 7: $\hat{\beta} = -0.17$, 95% CI [-1.40,1.06], $t(132) = -0.27$, $p = .791$). Interpreting these results, 5th graders improved 2.25 points out of 18 on average and 7th graders improved 1.94 points on average (Figure 4.4.3).

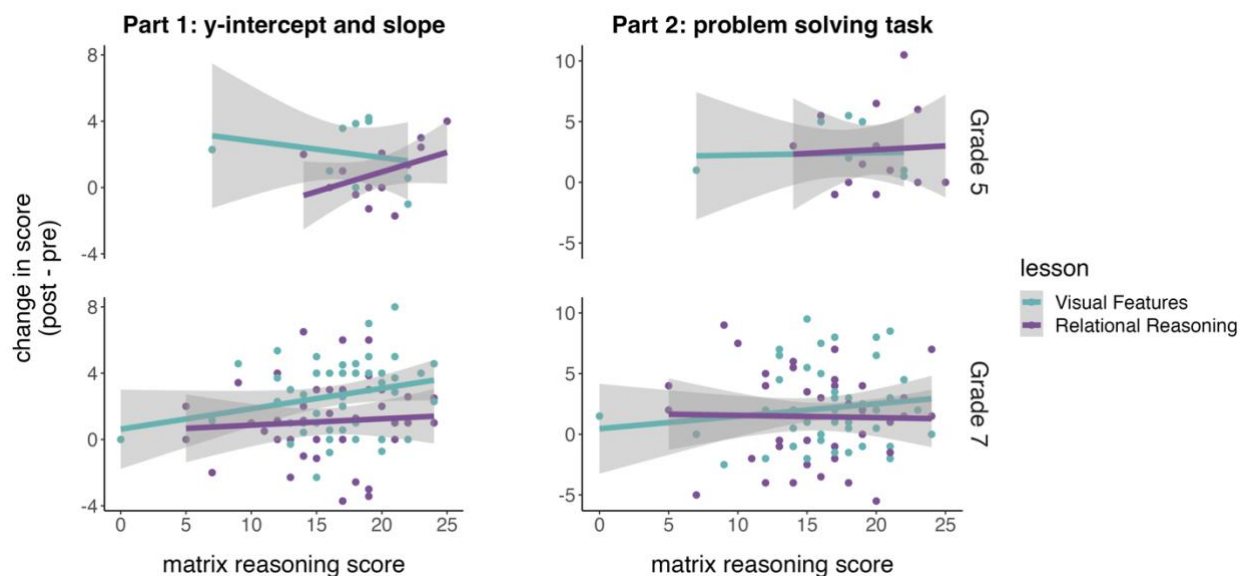


Figure 4.4.4. Relation between matrix reasoning and change in graph score by grade and lesson condition. In the left panel, the outcome measure is the difference between the part 1 score on the posttest minus pretest. In the right panel, the outcome measure is the difference between part 2 score on the posttest minus pretest. Points and lines are colored by lesson condition. The gray portion shows the 95% confidence interval around the linear model.

Matrix reasoning did not predict change in graph scores

Finally, I fit linear models to test for the effects of matrix reasoning on graph score, as well as for an interaction effect between matrix score and condition. Models predicting the change in part 1 score from matrix reasoning score, condition, and their interaction, revealed no effect of matrix reasoning for either grade (Grade 5: $b = -0.10$, 95% CI $[-0.39, 0.19]$, $t(19) = -0.72$, $p = .478$; Grade 7: $b = 0.12$, 95% CI $[-0.03, 0.28]$, $t(88) = 1.60$, $p = .113$), nor an interaction effect (Grade 5: $b = 0.34$, 95% CI $[-0.11, 0.79]$, $t(19) = 1.57$, $p = .134$; Grade 7: $b = -0.08$, 95% CI $[-0.30, 0.13]$, $t(88) = -0.77$, $p = .444$) (Figure 4.4.4, left panel). Similarly, models predicting change in part 2 score from the same predictors revealed the same result, no effect of matrix reasoning (Grade 5: $b = 0.02$, 95% CI $[-0.51, 0.55]$, $t(19) = 0.08$, $p = .940$; Grade 7: $b = 0.10$, 95% CI $[-0.12, 0.33]$, $t(88) = 0.90$, $p = .372$) nor an interaction between condition and matrix reasoning (Grade 5: $b = 0.04$, 95% CI $[-0.77, 0.85]$, $t(19) = 0.11$, $p = .914$; Grade 7: $b = -0.12$, 95% CI $[-0.45, 0.20]$, $t(88) = -0.74$, $p = .460$) (Figure 4.4.4, right panel). These results suggest that across conditions participants from both grades improved their understanding of y-intercept and slope and improved their problem solving with graphs regardless of their matrix reasoning score.

4.5 Discussion

The primary aims of the present chapter were to put forth the claim that relational reasoning is a foundational skill for graph comprehension, and to begin exploring this claim empirically. I proposed that relational reasoning—the domain-general cognitive capacity to map abstract relations between representations—allows individuals to map the visual features and patterns presented in graphs to their real-world referents and meaning. Specifically, I highlighted roles for

relational reasoning and the related concept of relational complexity in three main aspects of graph comprehension: in cognitively processing graphs, in the levels of difficulty with graph prompts, and in the complexity of the represented data. Throughout this discussion, I emphasized how relational reasoning could support graph comprehension and, on the flipside, how relational complexity could serve as a processing bottleneck, potentially limiting understanding. Next, I proposed various ways that relational reasoning could be scaffolded during graph learning to enhance comprehension.

In the second part of the chapter, I conducted a preliminary study to begin testing these recommendations empirically. This study also served as a proof-of-concept that graph lessons can be designed to scaffold relational reasoning. I used a pretest-posttest intervention design to investigate the benefits of emphasizing higher-order relational reasoning during a lesson on graphs. The study focused on graphs of linear functions, and I worked with 5th and 7th grade students because in these grades students start working more with graphs of continuous variables, making it an ideal time to examine line graph learning. There were two well-match lesson conditions that taught the concepts of y-intercept and slope: the visual features (VF) lesson focused on the visual features and patterns of these concepts and the higher-order relational reasoning (RR) lesson focused on mapping these concepts to their meaning in the context of the axis variables. The 5th grade math teacher shared that this lesson was likely her students' first formal exposure to the concepts of y-intercept and slope. Slope first appears in the Common Core Math standards in 7th grade, and y-intercept first appears in 8th grade in a standard that also mentions slope for just the second time. (Common Core State Standards Initiative, 2010). Though these concepts were likely review or elaboration for many of the 7th graders—the study was at the end of the school year and half of the 7th graders were taking the equivalent of 8th grade math—we did not observe evidence of ceiling effects at pretest and the 7th grade pretest scores were not different than the 5th grader scores, suggesting that the 7th graders were still gaining expertise with graphs. Further, one of the 7th grade teachers shared that her students had likely never seen a graph with three lines on it in the classroom (e.g., the pre/posttest graphs), and possibly not a graph with two lines either. Thus, parts of the lesson were novel for 5th graders and 7th graders alike. This preliminary study yielded three main findings.

First, I found that matrix reasoning, a common measure of relational reasoning, was related to initial graph comprehension, though the pattern of results was different between grades 5 and 7. Matrix reasoning score predicted the quality of student's explanations of the concepts y-intercept and slope at pretest for 5th graders but not 7th graders. On the other hand, this score predicted problem-solving with graphs at pretest for 7th graders but not 5th graders. These results support a link between relational reasoning and graph comprehension. To our knowledge, this is the first study test this relation. This pattern of results also suggests that there may be some developmental or content knowledge differences in the role that relational reasoning plays. Perhaps relational reasoning is a stronger predictor of graph concept knowledge when students are early on in their learning, which could explain why we found matrix reasoning was a predictor of concept understanding for 5th graders but not 7th graders. This pattern of results would be consistent with the idea that relational reasoning supports the initial learning phase, predicting mastery of new mathematical skills. Then, once students are more familiar with a concept,

relational reasoning comes online for problem solving and transferring concepts to new contexts, which could explain why we found matrix reasoning was a predictor of graph problem-solving for 7th graders but not 5th graders. However, it is possible that our 5th grade sample (which was roughly 1/5 of the size of our 7th grade sample) was underpowered to detect a relationship between matrix reasoning and graph problem solving given that the effect sizes were comparable between the two grades. That said, the 5th-grade sample was sufficiently large to detect a significant relationship between matrix reasoning and understanding of basic concepts, an effect size that was almost twice as large.

The second finding was that students from both grades across both conditions showed improved understanding of graph concepts as well as improved problem-solving from pretest to posttest. This result is particularly noteworthy considering the brevity of the lesson (about 40 minutes of class time total) and that many of the graphs were more complex than what students had been exposed to previously. Specifically, all the graphs in the practice block showed an interaction effect with two lines, meaning a third variable was necessary for reasoning about the relation between x and y . Though this type of graph was used facilitate comparison and engage higher-order relational reasoning, a secondary effect of using these graphs was that late elementary and middle school students were effectively practicing interpreting interactions, a skill that is not even included in the high school standards! Interestingly, 7th graders in the VF condition improved their explanations of the concepts of y -intercept and slope more than their counterparts in the RR condition. This finding runs counter to the initial prediction. However, it is possible that both lesson conditions could benefit students, albeit in different ways; this possibility provides a direction for future research.

Despite the emphasis of the VF lesson on the visual features and patterns in the graphs, there were three aspects of the lesson design that may have made it equally, or more, beneficial than the RR lesson. First, because identifying and encoding visual patterns is the first step of graph processing, directing students' attention to the relevant visual information on the graph could generally scaffold their understanding. On the other hand, focusing on mapping higher-order relations may be relatively useless if the student has not first identified the visual pattern to interpret. Second, the questions in the VF lesson included the new terminology more often than those in the RR lesson, which emphasized to a greater extent the context of the graph. For example, one question in the VF lesson asked, "which student has the greater y -intercept?" whereas the version of that question in the RR lesson requested that value in the context of the axis labels: "If Isaiah and Rosa decided not to study for the test at all, which student would score higher on the test?" Therefore, students in the VF condition may have gained more experience with the terminology, whereas students in the RR lesson may have gained more experience with the concepts in context, but may not have attached the term to the concept. Third, to match the content of the two lessons as closely as possible, both presented students with graphs displaying two lines, a graph type that is not typically introduced in these grades. These graphs were intentionally included in the lesson to elicit higher-order relational reasoning in the RR condition by encouraging comparison. For example, looking at two lines on a graph, students in the RR condition were asked to make a range-based comparison in the context of the axis labels, such as "Which plant would be taller if they were both watered less than 15mL a day?" In the VF

condition, the corresponding question encouraged point-based reasoning by asking students to compare two points on the graph, such as “Which plant has the higher y -value when $x = 10$?” Though the types of comparisons that students were asked to make differed by condition, making any comparisons at all on a graph would theoretically tax higher-order relational reasoning. Further, just because students in the VF condition were only asked to compare points does not mean that they did not also compare the structure of the lines and notice other similarities and differences.

Together, these three aspects of the lesson design suggest that the VF lesson may have engaged more higher-order relational reasoning than had initially been intended. Additionally, the RR lesson questions could have been too difficult for more novice learners, since additional lines and range-based reasoning increases the relational complexity. This question type could be a useful scaffold for some learners but overload others. Future work should investigate this potential tradeoff and whether and for whom showing two lines supports learning. Relatedly, a future study could also compare learning materials with two lines on one graph, versus comparing graphs side-by-side, versus no comparison. As described in more detail below, future studies should also include a contrast lesson that is more differentiated from the relational reasoning lesson and is more similar to typical classroom instruction, in order to better understand which features of these lessons are benefiting students.

The third result is that matrix reasoning did not predict the magnitude of change from pretest to posttest in students’ understanding or problem-solving for either condition. In other words, students with higher and lower relational reasoning scores benefited equally from the lessons. However, this study may have been underpowered to detect individual differences in students’ improvements, particularly because there could be an interaction between reasoning performance and lesson type. If such a finding were obtained in a larger sample, it would have important applications to pedagogy by shedding light on how to match a learner’s prior knowledge and skill set with a lesson that would benefit them the most. This question is particularly interesting because the direction of this interaction is not straightforward. For example, students with less graph proficiency or lower relational reasoning may benefit more from a lesson that focuses on visual features to help them attend to relevant visual relations, whereas more proficient graph users or students with higher relational reasoning may benefit more from a lesson that engages higher-order relational reasoning to deepen their understanding. Alternatively, students with lower relational reasoning may in fact benefit more from a relational reasoning-focused lesson to help them scaffold these skills that would be more difficult for them to engage without external support.

Limitations and future directions

The empirical preliminary study had three main limitations. First, the data were collected remotely in spring 2021, which was still during the Covid-19 pandemic. Though the 5th and 7th grade participants were in person at school with their teacher, their school year had been disrupted, and they were likely to have been behind in math, including graph knowledge. Thus, the study should be replicated to make sure the results hold under more typical learning conditions. Further, the pandemic made it difficult to recruit teachers to participate, and teachers

were less willing to give as much class time to the study because instructional time was often limited and unpredictably disrupted. This led to the large differences in sample size between 5th, 6th, and 7th grade, and may have resulted in our sample being underpowered to detect certain effects. It also led to the need for the study to be compressed into two days instead of the originally planned three days, thereby shortening the lessons and pre/posttest, and resulted in many 7th grade students not having enough time to finish the matrix reasoning task, which was the last task on the second day.

A second limitation was in the sensitivity of our measures of graph competency. For a preliminary study, it was beneficial to use open-ended questions to get a richer qualitative view into how students were answering questions and reasoning with graphs. Due to the timing constraints described above and the amount of time it takes students to answer open-ended questions, the pretest and posttest could not include many questions. Further, although we invested significant effort into developing and validating a rubric for coding the responses, in some cases it was still difficult to gauge students' comprehension, especially when their justifications were brief. Future studies could address these limitations by using rich multiple-choice questions constructed from the open-ended responses from this preliminary study. This would make it possible to include a larger number of pre- and post-test questions, and to identify both correct answers as well as patterns of misconceptions observed in the preliminary study, in a way that is easier to score and analyze. Measures could also be adapted from assessments of graph comprehension published after this preliminary study was run (e.g., Ciccione et al., 2023; Lloyd et al., 2023) or from assessment tools created for adults to make them more appropriate for children (e.g., Maltese et al., 2015). With additional class time, future studies would also be able to include a more comprehensive general measures of students' graph comprehension fluency, such as the Test of Graphing in Science (TOGS; McKenzie & Padilla, 1986), and could even examine change in scores as a result of the lesson.

A third limitation was that the study did not have an active or passive control lesson that did not involve graph comprehension. Both lessons had features that could be expected to help students improve, and indeed this is what we found. However, we cannot be sure that these improvements were not due to practice effects. Therefore, future studies should include an additional control group. Further, though the well-matched lesson design was beneficial for a preliminary study, future work should better differentiate the two experimental lesson conditions. For example, only the higher-order relational reasoning lesson should have graphs with two lines on them. Further, this lesson could more explicitly teach the deep relational structure of graphs, including highlighting the reasons one would choose to construct a graph in the first place.

Given that this was a preliminary study, there are many directions for future work. First, future research should extend beyond graphs of linear functions to investigate the effect of scaffolding relational reasoning during learning to interpret graphs of real-world data, which have additional complexities like variability and error that require more inferences to interpret. Future work should also focus on the role of relational reasoning in graph creation, not just interpreting graphs that have been created by someone else. Graph creation could be integrated as part of a full series of graph lessons that are designed to scaffold relational reasoning at each stage of learning

and practice. Though the present work focuses on graphs, this relational reasoning perspective can be extended to all types of data visualizations and diagrams, all of which represent non-spatial relations with spatial relations. Thus, a relational reasoning perspective has broad applications, and more research is needed to fully understand its benefits for effectively learning how to use and interpret these external representations.

Conclusion

Be it in the context of a K-12 math or science class, a graduate statistics course, or the homepage of the New York Times, graphs are a useful tool for discovering and communicating relations in data. However, their deep relational structure—both in terms of what and how information is represented—is not always apparent. The present work makes a strong theoretical case for the role of relational reasoning in graph comprehension and provides a productive framework for adding to our understanding of what makes this set of skills difficult as well as for designing new pedagogical approaches to address these obstacles. The preliminary empirical findings reported here support this view, and there are many future research directions. Given the importance of graph comprehension and the promise of this relational reasoning approach, I advocate for more psychology research on graph learning in children. This line of work has broad implications and the potential to help both children and adults make more sense of the sea of data around them.

General Discussion

Summary and theoretical implications

The present dissertation investigated the dynamics of relational reasoning as both a cognitive tool and bottleneck, and explored how offloading relations to external representations can help overcome cognitive limitations. In Chapter 1, I reviewed evidence that the protracted, immersive experience of formal schooling taxes, and therefore improves, general reasoning skills, such as relational reasoning. This review showed that in addition to reasoning supporting academic achievement and reasoning skills changing across developing, the experience of education itself affects reasoning and its development.

In Chapter 2, I showed that relational reasoning is a separable cognitive process from other domain-general processes that are related to and often co-occur with it, such as working memory. I found that although the executive functions of working memory and inhibitory control are robust predictors of fraction understanding, relational reasoning explained additional variance over and above these predictors. By connecting the executive function, relational reasoning, and math cognition literatures, this chapter established that relational reasoning should be considered a distinct core cognitive ability that uniquely contributes to academic achievement and laid the groundwork for the research presented in Chapter 4 on how this ability should be engaged to scaffold learning.

In Chapter 3, I demonstrated that offloading relations to external representations is part of a foundational cognitive toolkit and is separate from the regular use of visuospatial tools. I found that individuals spontaneously offloaded to-be-remembered relations to physical space on the table in front of them, including individuals who reported no formal schooling and were not literate. These results suggest that relational offloading is available as a cognitive resource for reducing relational demand even without the influences of formal schooling and using other formal visuospatial tools, such as writing. Dedre Gentner (2014) once said, “Space is the universal donor of relational thinking”, and though this sentiment is generally agreed upon, it has rarely been directly tested because relational reasoning and spatial cognition are often siloed areas of research. The results from this study provide direct evidence in support of Gentner’s claim. More broadly, this chapter establishes the importance of relational offloading for thinking, remembering, and reasoning and reveals the cognitive basis for the invention and use of visuospatial tools, laying the empirically groundwork for future studies examining when and how individuals offload relational demand to space.

Finally, in Chapter 4, I showed how relational reasoning can be scaffolded during instruction to support learning, using the case of graph comprehension. First, I proposed a relational reasoning perspective on graph comprehension, identifying various ways that this cognitive skill may support graph interpretation and how many of the comprehension difficulties can be reframed as cognitive bottlenecks due to relational complexity. Then, I implemented this approach and designed instructional materials that scaffolded relational reasoning to help students improve their understanding of important graph concepts. This preliminary study shows that helping students engage their relational reasoning by focus on patterns—both visual and conceptual—

and making comparisons can improve learning. It also served as a proof-of-concept that it is feasible to design lessons that directly scaffold relational reasoning, and points to many future directions for research and pedagogy.

Pedagogical implications

Based on the findings in this dissertation, I propose that applying the lens of relational reasoning can help improve STEM education in three significant ways. First, it can help identify what STEM content students may find particularly difficult, and therefore what topics may require more attention during instruction. Given that relational complexity is a rate-limiting factor, STEM content can be analyzed for areas that may be particularly complex, as was done in Chapter 4 for graph comprehension. For example, density, which is often considered one of the most difficult science concepts taught in middle school (e.g., C. L. Smith et al., 1997), is also relationally complex since it is the relation between mass and volume. Indeed, one common misconception is that students confuse density for weight, a unary variable, and do not consider density's inherent relational nature (C. L. Smith et al., 1997). Statistics is another ripe area for this type of relational analysis since the bulk of statistics relies on proportions and focuses on inferences, and it is often perceived as being extremely difficult by students (Son et al., 2021).

Second, the lens of relational reasoning can inform the design of pedagogical approaches for helping students overcome such learning obstacles, as was exemplified in Chapter 4. This lens can inform *how* to teach those particularly relationally demanding concepts and skills and scaffold relational reasoning, such as what external representations or visuospatial tools may be useful, how to break concepts down into more manageable smaller parts, what relational features and patterns to draw students' attention to, and how comparison can be utilized to engage relational reasoning. These pedagogical approaches can help make relations that may be opaque to novices more explicit and overt during the learning process. The graph lessons in Chapter 4 provide concrete examples of this approach. Future work should apply this lens more broadly, such as to density and concepts in statistics.

Finally, a relational reasoning approach can also inform the design of the tools themselves, such as better data visualizations or diagrams. Though there is already literature on cognitive science approaches to the design of these visuospatial tools (e.g., Franconeri et al., 2021; Hegarty, 2011), these studies and recommendations have yet to incorporate a relational reasoning perspective. Work that has begun to apply this approach, such as by structurally aligning components of a display to facilitate visual comparison (Matlen et al., 2020), has been successful in improving understanding, but more work is needed.

The research in this dissertation along with these three areas that stand to benefit from applying the lens of relational reasoning have practical implications for teaching at all levels, from elementary to graduate education. Instructors should highlight that visuospatial tools, such as graphs and diagrams, are human-invented tools and should explain what they were invented for, how they achieve this purpose, and when and why they are used. This includes making their relational structures explicit, concrete, visible, and meaningful. Though this meta-knowledge may be obvious to individuals with a great deal of experience and expertise, it is often not obvious to

novices, nor is it typically communicated to them, and it could be foundational for their robust, flexible, and transferrable understanding of the tools themselves as well as the concepts they learn using the tools. Further, students need practice mapping relations and successfully completing the most complex steps. Students do not learn how to do these steps simply by seeing an external representation or seeing someone else reason with it, they must practice themselves, including creating diagrams and graphs.

Conclusion

This dissertation work draws on and contributes to many literatures due its interdisciplinary nature. Specifically, I bridge the psychology literatures of relational reasoning, graph comprehension, visual reasoning, scientific reasoning, cognitive offloading, spatial cognition, math cognition, and executive functions. In addition, I have drawn inspiration from education research in the subfields of math, science, data science, and statistics education, as well as the visualization literature from computer science. Taken together, this work simultaneously contributes to our understanding of the role of relational reasoning in higher-order cognition, relational offloading to external representations, and to applications of relational reasoning for improving STEM education. In particular, my research provides a generative framework for identifying the STEM content that students may find especially difficult, as well as for informing the design of pedagogical approaches for helping students overcome these obstacles.

References

- Abreu-Mendoza, R. A., Chamorro, Y., Garcia-Barrera, M. A., & Matute, E. (2018). The contributions of executive functions to mathematical learning difficulties and mathematical talent during adolescence. *PLoS ONE*, *13*(12), 0209267. <https://doi.org/10.1371/journal.pone.0209267>
- Abreu-Mendoza, R. A., Coulanges, L., Ali, K., Powell, A. B., & Rosenberg-Lee, M. (2020). Children's discrete proportional reasoning is related to inhibitory control and enhanced by priming continuous representations. *Journal of Experimental Child Psychology*, *199*, 104931. <https://doi.org/10.1016/j.jecp.2020.104931>
- Alexander, P. A. (2016). Relational thinking and relational reasoning: Harnessing the power of patterning. *Npj Science of Learning*, *1*(1), Article 1. <https://doi.org/10.1038/npjscilearn.2016.4>
- Alexander, P. A. (2017). Relational Reasoning in STEM Domains: A Foundation for Academic Development. *Educational Psychology Review*, *29*(1), 1–10. <https://doi.org/10.1007/s10648-016-9383-1>
- Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., Langen, J., & Kievit, R. A. (2021). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, *4*, 63. <https://doi.org/10.12688/wellcomeopenres.15191.2>
- Andrews, G., & Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology*, *45*(2), 153–219. [https://doi.org/10.1016/S0010-0285\(02\)00002-6](https://doi.org/10.1016/S0010-0285(02)00002-6)
- Armitage, K. L., Bulley, A., & Redshaw, J. (2020). Developmental origins of cognitive offloading. *Proceedings of the Royal Society B*, *287*(1928), 20192927–20192927. <https://doi.org/10.1098/RSPB.2019.2927>
- Armitage, K. L., & Redshaw, J. (2022). Children boost their cognitive performance with a novel offloading technique. *Child Development*, *93*(1), 25–38. <https://doi.org/10.1111/cdev.13664>
- Armitage, K. L., Taylor, A. H., Suddendorf, T., & Redshaw, J. (2022). Young children spontaneously devise an optimal external solution to a cognitive problem. *Developmental Science*, *25*(3). <https://doi.org/10.1111/DESC.13204>
- Atit, K., Power, J. R., Veurink, N., Uttal, D. H., Sorby, S., Panther, G., Msall, C., Fiorella, L., & Carr, M. (2020). Examining the role of spatial skills and mathematics motivation on middle school mathematics achievement. *International Journal of STEM Education*, *7*(1), 38. <https://doi.org/10.1186/s40594-020-00234-3>
- Atit, K., Uttal, D. H., & Stieff, M. (2020). Situating space: Using a discipline-focused lens to examine spatial thinking skills. *Cognitive Research: Principles and Implications*, *5*(1), 1–16. <https://doi.org/10.1186/S41235-020-00210-Z/METRICS>
- Augue, B. (2017). *gridExtra: Miscellaneous functions for "Grid" graphics* [Manual]. <https://CRAN.R-project.org/package=gridExtra>
- Aust, F., & Barth, M. (2022). *papaja: Prepare American Psychological Association Journal Articles with R Markdown*. <https://github.com/crsh/papaja>
- Avgerinou, V. A., & Tolmie, A. (2019). Inhibition and cognitive load in fractions and decimals. *The British Journal of Educational Psychology*. <https://doi.org/10.1111/bjep.12321>

- Bailey, D. H., Siegler, R. S., & Geary, D. C. (2014). Early predictors of middle school fraction knowledge. *Developmental Science, 17*(5), 775–785. <https://doi.org/10.1111/desc.12155>
- Balchin, W. G. V., & Coleman, A. M. (1966). Graphicacy should be the fourth ace in the pack. *Cartographica: The International Journal for Geographic Information and Geovisualization, 3*(1), 23–28. <https://doi.org/10.3138/C7Q0-MM01-6161-7315>
- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637. <https://doi.org/10.1037/0033-2909.128.4.612>
- Bascandziev, I., Tardiff, N., Zaitchik, D., & Carey, S. (2018). The role of domain-general cognitive resources in children’s construction of a vitalist theory of biology. *Cognitive Psychology, 104*, 1–28. <https://doi.org/10.1016/j.cogpsych.2018.03.002>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2023). *lme4: Linear Mixed-Effects Models using Eigen and S4*. <https://github.com/lme4/lme4/>
- Bauer, M. I., & Johnson-Laird, P. N. (1993). How Diagrams Can Improve Reasoning. *Psychological Science, 4*(6), 372–378. <https://doi.org/10.1111/j.1467-9280.1993.tb00584.x>
- Bergen, B. K., & Lau, T. T. C. (2012). Writing Direction Affects How People Map Space Onto Time. *Frontiers in Psychology, 3*. <https://doi.org/10.3389/fpsyg.2012.00109>
- Bertin, J. (1983). *Semiology of Graphics: Diagrams, Networks, Maps*. UMI Research Press.
- Bertin, J. (2001). Matrix theory of graphics. *Information Design Journal, 10*(1), 5–19.
- Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and Individual Differences, 21*(4), 327–336. <https://doi.org/10.1016/j.lindif.2011.01.007>
- Blackwell, K. A., Chatham, C. H., Wiseheart, M., & Munakata, Y. (2014). A developmental window into trade-offs in executive function: The case of task switching versus response inhibition in 6-year-olds. *Neuropsychologia, 62*, 356–364. <https://doi.org/10/f6mn8f>
- Bonato, M., Fabbri, S., Umiltà, C., & Zorzi, M. (2007). The mental representation of numerical fractions: Real or integer? *Journal of Experimental Psychology: Human Perception and Performance, 33*(6), 1410–1419. <https://doi.org/10.1037/0096-1523.33.6.1410>
- Bonesso, S., Bruni, E., & Gerli, F. (2020). How Big Data Creates New Job Opportunities: Skill Profiles of Emerging Professional Roles. In S. Bonesso, E. Bruni, & F. Gerli (Eds.), *Behavioral Competencies of Digital Professionals: Understanding the Role of Emotional Intelligence* (pp. 21–39). Springer International Publishing. https://doi.org/10.1007/978-3-030-33578-6_2
- Boote, S. K., & Boote, D. N. (2017). Leaping from Discrete to Continuous Independent Variables: Sixth Graders’ Science Line Graph Interpretations. *The Elementary School Journal, 117*(3), 455–484. <https://doi.org/10.1086/690204>

- Börner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences*, *116*(6), 1857–1864. <https://doi.org/10.1073/pnas.1807180116>
- Börner, K., Maltese, A., Balliet, R. N., & Heimlich, J. (2016). Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Information Visualization*, *15*(3), 198–213. <https://doi.org/10.1177/1473871615594652>
- Boroditsky, L., & Gaby, A. (2010). Remembrances of Times East: Absolute Spatial Representations of Time in an Australian Aboriginal Community. *Psychological Science*, *21*(11), 1635–1639. <https://doi.org/10.1177/0956797610386621>
- Braithwaite, D. W., Leib, E. R., Siegler, R. S., & McMullen, J. (2019). Individual differences in fraction arithmetic learning. *Cognitive Psychology*, *112*, 81–98. <https://doi.org/10.1016/j.cogpsych.2019.04.002>
- Braithwaite, D. W., & Siegler, R. S. (2018). Developmental changes in the whole number bias. *Developmental Science*, *21*(2), e12541. <https://doi.org/10.1111/desc.12541>
- Brand, T. van den. (2023). *ggh4x: Hacks for ggplot2*. <https://CRAN.R-project.org/package=ggh4x>
- Bransford, J. D., & Schwartz, D. L. (1999). Rethinking Transfer: A Simple Proposal With Multiple Implications. *Review of Research in Education*, *24*(1), 61–100. <https://doi.org/10.3102/0091732X024001061>
- Brich, I. R., Bause, I. M., Hesse, F. W., & Wesslein, A.-K. (2019). Working memory affine technological support functions improve decision performance. *Computers in Human Behavior*, *92*, 238–249. <https://doi.org/10.1016/j.chb.2018.11.014>
- Brod, G., Hasselhorn, M., & Bunge, S. A. (2018). When generating a prediction boosts learning: The element of surprise. *Learning and Instruction*, *55*, 22–31.
- Brookman-Byrne, A., Mareschal, D., Tolmie, A. K., & Dumontheil, I. (2018). Inhibitory control and counterintuitive science and maths reasoning in adolescence. *PLoS ONE*, *13*(6). <https://doi.org/10.1371/journal.pone.0198973>
- Bugden, S., & Ansari, D. (2011). Individual differences in children’s mathematical competence are related to the intentional but not automatic processing of Arabic numerals. *Cognition*, *118*(1), 32–44. <https://doi.org/10.1016/j.cognition.2010.09.005>
- Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives*, *8*(1), 36–41. <https://doi.org/10.1111/cdep.12059>
- Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children’s mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology*, *19*(3), 273–293.
- Bulley, A., McCarthy, T., Gilbert, S. J., Suddendorf, T., & Redshaw, J. (2020). Children Devise and Selectively Use Tools to Offload Cognition. *Current Biology*, *30*(17), 3457–3464.e3. <https://doi.org/10.1016/J.CUB.2020.06.035>
- Bunge, S. A., Dudukovic, N. M., Thomason, M. E., Vaidya, C. J., & Gabrieli, J. D. (2002). Immature frontal lobe contributions to cognitive control in children: Evidence from fMRI. *Neuron*, *33*(2), 301–311.
- Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical Reasoning and Prefrontal Cortex: Evidence for Separable Retrieval and Integration Mechanisms. *Cerebral Cortex*, *15*(3), 239–249. <https://doi.org/10.1093/cercor/bhh126>

- Bunge, S. A., & Zelazo, P. D. (2006). A brain-based account of the development of rule use in childhood. *Current Directions in Psychological Science*, *15*(3), 118–121. <https://doi.org/10.1111/j.0963-7214.2006.00419.x>
- Burnyeat, M. F. (2000). Plato on why mathematics is good for the soul. *Proceedings of the British Academy*, *103*, 1–81.
- Cahan, S., & Cohen, N. (1989). Age versus schooling effects on intelligence development. *Child Development*, *60*(5), 1239–1249. <https://doi.org/10.1111/j.1467-8624.1989.tb03554.x>
- Carey, S. (2011). Precis of “The Origin of Concepts.” *The Behavioral and brain sciences*, *34*(3), 113–124 124–162. <https://doi.org/10.1017/S0140525X10000919>
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404–431. <https://doi.org/10.1037/0033-295X.97.3.404>
- Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*(2), 75–100. <https://doi.org/10.1037/1076-898X.4.2.75>
- Carpenter, T. P., Fennema, E., & Franke, M. L. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *The Elementary School Journal*, *97*(1), 3–20. <https://doi.org/10.1086/461846>
- Cattell, R. B. (1940). A culture-free intelligence test I. *Journal of Educational Psychology*, *31*, 161–179.
- Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin*, *40*(3), 153–193. <https://doi.org/10.1037/h0059973>
- Cattell, R. B. (1987a). *Intelligence: Its structure, growth, and action*. North-Holland.
- Cattell, R. B. (1987b). The Discovery of Fluid and Crystallized General Intelligence. In *Intelligence: Its Structure, Growth and Action* (pp. 87–120). Elsevier Science Publishers.
- Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology*, *27*(5), 703–722. <https://doi.org/10.1037/0012-1649.27.5.703>
- Cetron, J. S., Connolly, A. C., Diamond, S. G., May, V. V., Haxby, J. V., & Kraemer, D. J. M. (2019). Decoding individual differences in STEM learning from functional MRI data. *Nature Communications*, *10*(1). <https://doi.org/10.1038/s41467-019-10053-y>
- Chapman, S. B., & Mudar, R. A. (2014). Enhancement of cognitive and neural functions through complex reasoning training: Evidence from normal and clinical populations. *Frontiers in Systems Neuroscience*, *8*. <https://doi.org/10.3389/fnsys.2014.00069>
- Chen, Z., Honomichl, R., Kennedy, D., & Tan, E. (2016). Aiming to complete the matrix: Eye-movement analysis of processing strategies in children’s relational thinking. *Developmental Psychology*, *52*(6), 867–878. <https://doi.org/10.1037/dev0000113>
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121–152.
- Chi, M. T. H., & VanLehn, K. A. (2012). Seeing deep structure from the interactions of surface features. *Educational Psychologist*, *47*(3), 177–188. <https://doi.org/10.1080/00461520.2012.695709>
- Christie, S., & Gentner, D. (2014). Language Helps Children Succeed on a Classic Analogy Task. *Cognitive Science*, *38*(2), 383–397. <https://doi.org/10.1111/cogs.12099>

- Christoff, K., Ream, J. M., Geddes, L. P. T., & Gabrieli, J. D. E. (2003). Evaluating self-generated information: Anterior prefrontal contributions to human cognition. *Behavioral Neuroscience*, *117*(6), 1161–1168. <https://doi.org/10.1037/0735-7044.117.6.1161>
- Ciccione, L., Sablé-Meyer, M., Boissin, E., Josserand, M., Potier-Watkins, C., Caparos, S., & Dehaene, S. (2023). Trend judgment as a perceptual building block of graphicacy and mathematics, across age, education, and culture. *Scientific Reports*, *13*(1), Article 1. <https://doi.org/10.1038/s41598-023-37172-3>
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19. <https://doi.org/10.1093/ANALYS/58.1.7>
- Clarke, D. M., & Roche, A. (2009). Students' fraction comparison strategies as a window into robust understanding and possible pointers for instruction. *Educational Studies in Mathematics*, *72*(1), 127–138. <https://doi.org/10.1007/s10649-009-9198-9>
- Clarke, E., & Sherrill-Mix, S. (2017). *ggbeeswarm: Categorical scatter (violin point) plots* [Manual]. <https://CRAN.R-project.org/package=ggbeeswarm>
- Collins, A., & Ferguson, W. (1993). Epistemic forms and Epistemic Games: Structures and Strategies to Guide Inquiry. *Educational Psychologist*, *28*(1), 25–42. https://doi.org/10.1207/s15326985ep2801_3
- Common Core State Standards Initiative. (2010). *Common Core State Standards for Mathematics*. National Governors Association Center for Best Practices and the Council of Chief State School Officers. https://learning.ccsso.org/wp-content/uploads/2022/11/Math_Standards1.pdf
- Constantinidis, C., & Luna, B. (2019). Neural substrates of inhibitory control maturation in adolescence. *Trends in Neurosciences*, *42*(9), 604–616. <https://doi.org/10.1016/j.tins.2019.07.004>
- Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552. <https://doi.org/10.1016/j.tics.2003.10.005>
- Cooperrider, K., Marghetis, T., & Núñez, R. (2017). Where Does the Ordered Line Come From? Evidence From a Culture of Papua New Guinea. *Psychological Science*, *28*(5), 599–608. https://doi.org/10.1177/0956797617691548/ASSET/IMAGES/LARGE/10.1177_0956797617691548-FIG3.JPEG
- Corsi, P. (1972). *Human memory and the medial temporal region of the brain*.
- Coulanges, L., Abreu-Mendoza, R. A., Varma, S., Uncapher, M. R., Gazzaley, A., Anguera, J., & Rosenberg-Lee, M. (2021). Linking inhibitory control to math achievement via comparison of conflicting decimal numbers. *Cognition*, *214*, 104767. <https://doi.org/10.1016/j.cognition.2021.104767>
- Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, *26*(2), 197–223. <https://doi.org/10.1007/s10648-013-9246-y>
- Cragg, L., & Gilmore, C. (2014). Skills underlying mathematics: The role of executive function in the development of mathematics proficiency. *Trends in Neuroscience and Education*, *1*–6. <https://doi.org/10.1016/j.tine.2013.12.001>

- Crone, E. A., Wendelken, C., Van Leijenhorst, L., Honomichl, R. D., Christoff, K., & Bunge, S. A. (2009). Neurocognitive development of relational reasoning. *Developmental Science*, *12*(1), 55–66. <https://doi.org/10.1111/j.1467-7687.2008.00743.x>
- Curcio, F. R. (1987). Comprehension of Mathematical Relationships Expressed in Graphs. *Journal for Research in Mathematics Education*, *18*(5), 382–393. <https://doi.org/10.5951/jresematheduc.18.5.0382>
- Data Science and Literacy Act of 2023, H.R.1050, 118th Congress (2023). <https://www.congress.gov/bill/118th-congress/house-bill/1050>
- Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in western and Amazonian indigene cultures. *Science*, *320*(5880), 1217–1220. https://doi.org/10.1126/SCIENCE.1156540/SUPPL_FILE/DEHAENE.SOM.PDF
- DeWolf, M., Bassok, M., & Holyoak, K. J. (2015). Conceptual structure and the procedural affordances of rational numbers: Relational reasoning with fractions and decimals. *Journal of Experimental Psychology: General*, *144*(1), 127–150. <https://doi.org/10.1037/xge0000034>
- DeWolf, M., & Vosniadou, S. (2015). The representation of fraction magnitudes and the whole number bias reconsidered. *Learning and Instruction*, *37*, 39–49. <https://doi.org/10.1016/j.learninstruc.2014.07.002>
- Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. <https://doi.org/10.1146/annurev-psych-113011-143750>
- Diezmann, C. M. (2000). Making sense with diagrams: Students' difficulties with feature-similar problems. *23rd Annual Conference of Mathematics Education Research Group of Australasia*, 228–234. <https://eprints.qut.edu.au/1501/>
- Dimara, E., Franconeri, S., Plaisant, C., Bezerianos, A., & Dragicevic, P. (2020). A Task-Based Taxonomy of Cognitive Biases for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *26*(2), 1413–1432. <https://doi.org/10.1109/TVCG.2018.2872577>
- DiSessa, A. A., Hammer, D., Sherin, B., & Kolpakowski, T. (1991). Inventing graphing: Meta-representational expertise in children. *The Journal of Mathematical Behavior*, *10*(2), 117–160.
- Dumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–43. <https://doi.org/10.1037/0033-295X.115.1.1>
- Drozda, Z. (2023). *Data Science Is Vital to Student Success. So Why Are Outcomes Going Down?* (p. 10). Data Science 4 Everyone.
- Dumas, D. (2017). Relational Reasoning in Science, Medicine, and Engineering. *Educational Psychology Review*, *29*(1), 73–95. <https://doi.org/10.1007/s10648-016-9370-6>
- Dumas, D., Alexander, P. A., & Grossnickle, E. M. (2013). Relational Reasoning and Its Manifestations in the Educational Context: A Systematic Review of the Literature. *Educational Psychology Review*, *25*(3), 391–427. <https://doi.org/10.1007/s10648-013-9224-4>
- Dumas, D., & Dong, Y. (2022). Relational Reasoning and Thinking: Theory, Measurement, and Empirical Findings. In *Relational Reasoning and Thinking: Theory, Measurement, and Empirical Findings*. Routledge. <https://doi.org/10.4324/9781138609877-REE179-1>

- Dumontheil, I., Houlton, R., Christoff, K., & Blakemore, S.-J. (2010). Development of relational reasoning during adolescence. *Developmental Science*, *13*(6), F15–F24. <https://doi.org/10/dqppqh>
- Duncan, J., Schramm, M., Thompson, R., & Dumontheil, I. (2012). Task rules, working memory, and fluid intelligence. *Psychonomic Bulletin & Review*, *19*(5), 864–870. <https://doi.org/10.3758/s13423-012-0225-y>
- Dunn, T. L., & Risko, E. F. (2016). Toward a Metacognitive Account of Cognitive Offloading. *Cognitive Science*, *40*(5), 1080–1127. <https://doi.org/10.1111/cogs.12273>
- Eckstein, M. K., Guerra-Carrillo, B., Miller Singley, A. T., & Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, *25*, 69–91. <https://doi.org/10.1016/J.DCN.2016.11.001>
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309–331. <https://doi.org/10.1037/0096-3445.128.3.309>
- Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. <https://doi.org/10.3758/BF03203267>
- Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*(1), 1–63. [https://doi.org/10.1016/0004-3702\(89\)90077-5](https://doi.org/10.1016/0004-3702(89)90077-5)
- Farrell Pagulayan, K., Busch, R. M., Medina, K. L., Bartok, J. A., & Krikorian, R. (2006). Developmental normative data for the corsi block-tapping task. *Journal of Clinical and Experimental Neuropsychology*, *28*(6), 1043–1052. <https://doi.org/10.1080/13803390500350977>
- Fazio, L. K., DeWolf, M., & Siegler, R. S. (2016). Strategy use and strategy choice in fraction magnitude comparison. *Journal of Experimental Psychology-Learning Memory and Cognition*, *42*(1), 1–16. <https://doi.org/10.1037/xlm0000153>
- Fedden, S., & Boroditsky, L. (2012). Spatialization of Time in Mian. *Frontiers in Psychology*, *3*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00485>
- Ferrer, E., & McArdle, J. J. (2004). An experimental analysis of dynamic hypotheses about cognitive abilities and achievement from childhood to early adulthood. *Developmental Psychology*, *40*(6), 935–952. <https://doi.org/10.1037/0012-1649.40.6.935>
- Ferrer, E., McArdle, J. J., Shaywitz, B. A., Holahan, J. M., Marchione, K., & Shaywitz, S. E. (2007). Longitudinal models of developmental dynamics between reading and cognition from childhood to adolescence. *Developmental Psychology*, *43*(6), 1460–1473. <https://doi.org/10.1037/0012-1649.43.6.1460>
- Floyd, S. (2016). Modally Hybrid Grammar? Celestial Pointing for Time-of-Day Reference in Nheengatú. *Language*, *92*(1), 31–64.
- Fox, A. R. (2023). Theories and Models in Graph Comprehension. In D. Albers Szafir, R. Borgo, M. Chen, D. J. Edwards, B. Fisher, & L. Padilla (Eds.), *Visualization Psychology* (pp. 39–64). Springer International Publishing. https://doi.org/10.1007/978-3-031-34738-2_2

- Fox, J., Venables, B., Damico, A., & Salverda, A. P. (2021). *english: Translate Integers into English*. <https://CRAN.R-project.org/package=english>
- Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The Science of Visual Data Communication: What Works. *Psychological Science in the Public Interest*, 22(3), 110–161. <https://doi.org/10.1177/15291006211051956>
- Freedman, E. G., & Shah, P. (2002). Toward a Model of Knowledge-Based Graph Comprehension. In M. Hegarty, B. Meyer, & N. H. Narayanan (Eds.), *Diagrammatic Representation and Inference* (pp. 18–30). Springer. https://doi.org/10.1007/3-540-46037-3_3
- Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, 133(1), 101–135. <https://doi.org/10.1037/0096-3445.133.1.101>
- Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., Defries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science*, 17(2), 172–179. <https://doi.org/10.1111/j.1467-9280.2006.01681.x>
- Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making Sense of Graphs: Critical Factors Influencing Comprehension and Instructional Implications. *Journal for Research in Mathematics Education*, 32(2), 124–158. <https://doi.org/10.2307/749671>
- Friso-van Den Bos, I., Van Der Ven, S. H. G., Kroesbergen, E. H., & Van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review*, 10, 29–44. <https://doi.org/10.1016/j.edurev.2013.05.003>
- Fry, A. F., & Hale, S. (2000). Relationships among processing speed, working memory, and fluid intelligence in children. *Biological Psychology*, 54(1–3), 1–34. <https://doi.org/10/dpnh8s>
- Fu, X., Li, X., Xu, P., & Zeng, J. (2020). Inhibiting the Whole Number Bias in a Fraction Comparison Task: An Event-Related Potential Study. *Psychology Research and Behavior Management, Volume 13*, 245–255. <https://doi.org/10.2147/PRBM.S240263>
- Fuchs, L. S., Compton, D. L., Fuchs, D., Powell, S. R., Schumacher, R. F., Hamlett, C. L., Vernier, E., Namkung, J. M., & Vukovic, R. K. (2012). Contributions of domain-general cognitive resources and different forms of arithmetic development to pre-algebraic knowledge. *Developmental Psychology*, 48(5), 1315–1326. <https://doi.org/10.1037/a0027475>
- Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., Schatschneider, C., & Fletcher, J. M. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, 98(1), 29–43. <https://doi.org/10.1037/0022-0663.98.1.29>
- Fuhrman, O., & Boroditsky, L. (2010). Cross-Cultural Differences in Mental Representations of Time: Evidence From an Implicit Nonlinguistic Task. *Cognitive Science*, 34(8), 1430–1451. <https://doi.org/10.1111/j.1551-6709.2010.01105.x>
- Gabrieli, J. D. E. (2016). The promise of educational neuroscience: Comment on bowers (2016). *Psychological Review*, 123(5), 613–619. <https://doi.org/10.1037/rev0000034>
- Gamino, J. F., Motes, M. M., Riddle, R., Lyon, G. R., Spence, J. S., & Chapman, S. B. (2014). Enhancing inferential abilities in adolescence: New hope for students in poverty. *Frontiers in Human Neuroscience*, 8. <https://www.frontiersin.org/articles/10.3389/fnhum.2014.00924>

- Gardiner, A., Aasheim, C., Rutner, P., & Williams, S. (2018). Skill Requirements in Big Data: A Content Analysis of Job Advertisements. *Journal of Computer Information Systems*, 58(4), 374–384. <https://doi.org/10.1080/08874417.2017.1289354>
- Gattis, M. (2002). Structure mapping in spatial reasoning. *Cognitive Development*, 17(2), 1157–1183. [https://doi.org/10.1016/S0885-2014\(02\)00095-3](https://doi.org/10.1016/S0885-2014(02)00095-3)
- Gattis, M., & Holyoak, K. J. (1996). Mapping conceptual to spatial relations in visual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(1), 231–239. <https://doi.org/10.1037/0278-7393.22.1.231>
- Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: A 5-year longitudinal study. *Developmental Psychology*, 47(6), 1539–1552. <https://doi.org/10.1037/a0025510>
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155–170. [https://doi.org/10.1016/S0364-0213\(83\)80009-3](https://doi.org/10.1016/S0364-0213(83)80009-3)
- Gentner, D. (1988). Metaphor as Structure Mapping: The Relational Shift. *Child Development*, 59(1), 47–59. <https://doi.org/10.2307/1130388>
- Gentner, D. (2003). Why We're So Smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in Mind: Advances in the Study of Language and Thought* (pp. 195–235). MIT Press.
- Gentner, D. (2014, September). *Comparison and relational language in the development of relational categories*. [Public Opening Keynote Lecture]. 10th International Symposium of Cognition, Logic and Communication, Riga, Latvia.
- Gentner, D., Shao, R., Simms, N., & Hespos, S. (2021). Learning same and different relations: Cross-species comparisons. *Current Opinion in Behavioral Sciences*, 37, 84–89. <https://doi.org/10/ghtp84>
- Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12(3), 306–355. [https://doi.org/10.1016/0010-0285\(80\)90013-4](https://doi.org/10.1016/0010-0285(80)90013-4)
- Gilbert, S. J., Boldt, A., Sachdeva, C., Scarampi, C., & Tsai, P.-C. (2023). Outsourcing Memory to External Tools: A Review of 'Intention Offloading.' *Psychonomic Bulletin & Review*, 30(1), 60–76. <https://doi.org/10.3758/s13423-022-02139-4>
- Glazer, N. (2011). Challenges with graph interpretation: A review of the literature. *Studies in Science Education*, 47(2), 183–210. <https://doi.org/10.1080/03057267.2011.605307>
- Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, 142(7), 729–757. <https://doi.org/10.1037/bul0000043>
- Gómez, D. M., & Dartnell, P. (2018). Middle schoolers' biases and strategies in a fraction comparison task. *International Journal of Science and Mathematics Education*, 17(6), 1233–1250. <https://doi.org/10.1007/s10763-018-9913-z>
- Gómez, D. M., Jiménez, A., Bobadilla, R., Reyes, C., & Dartnell, P. (2015). The effect of inhibitory control on general mathematics achievement and fraction comparison in middle school children. *ZDM : The International Journal on Mathematics Education*, 47(5), 801–811. <https://doi.org/10.1007/s11858-015-0685-4>
- Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, 4(5), 178–186. [https://doi.org/10.1016/s1364-6613\(00\)01467-4](https://doi.org/10.1016/s1364-6613(00)01467-4)

- González-Forte, J. M., Fernández, C., & Dooren, W. (2018). Gap and congruency effect in fraction comparison. *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education*.
- González-Forte, J. M., Fernández, C., Hoof, J., & Dooren, W. (2020). Various ways to determine rational number size: An exploration across primary and secondary education. *European Journal of Psychology of Education, 35*(3), 549–565. <https://doi.org/10.1007/s10212-019-00440-w>
- Goswami, U. (2001). Analogical reasoning in children. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The Analogical Mind: Perspectives from Cognitive Science* (pp. 437–470). The MIT Press. <https://doi.org/10.7551/mitpress/1251.001.0001>
- Green, C. T., Bunge, S. A., Chiongbian, V. B., Barrow, M., & Ferrer, E. (2017). Fluid reasoning predicts future mathematical performance among children and adolescents. *Journal of Experimental Child Psychology, 157*, 125–143. <https://doi.org/10.1016/j.jecp.2016.12.005>
- Grinschgl, S., & Neubauer, A. C. (2022). Supporting Cognition With Modern Technology: Distributed Cognition Today and in an AI-Enhanced Future. *Frontiers in Artificial Intelligence, 5*. <https://www.frontiersin.org/articles/10.3389/frai.2022.908261>
- Grossnickle, E. M., Dumas, D., Alexander, P. A., & Baggetta, P. (2016). Individual differences in the process of relational reasoning. *Learning and Instruction, 42*, 141–159. <https://doi.org/10.1016/j.learninstruc.2016.01.013>
- Guerra-Carrillo, B., & Bunge, S. A. (2018). Eye gaze patterns reveal how reasoning skills improve with experience. *Npj Science of Learning, 3*(1), Article 1. <https://doi.org/10.1038/s41539-018-0035-8>
- Guerra-Carrillo, B., Katovich, K., & Bunge, S. A. (2017). Does higher education hone cognitive functioning and learning efficacy? Findings from a large and diverse sample. *PLOS ONE, 12*(8). <https://doi.org/10.1371/journal.pone.0182276>
- Halford, G. S., Andrews, G., Wilson, W. H., & Phillips, S. (2012). Computational models of relational processes in cognitive development. *Cognitive Development, 27*(4), 481–499. <https://doi.org/10.1016/j.cogdev.2012.08.003>
- Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences, 21*(6), 803–831. <https://doi.org/10.1017/S0140525X98001769>
- Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences, 14*(11), 497–505. <https://doi.org/10.1016/j.tics.2010.08.005>
- Halpern, D. F. (2001). Assessing the effectiveness of critical thinking instruction. *The Journal of General Education, 50*(4), 270–286. <https://doi.org/10.1353/jge.2001.0024>
- Hebb, D. O. (1942). The Effect of Early and Late Brain Injury upon Test Scores, and the Nature of Normal Adult Intelligence. *Proceedings of the American Philosophical Society, 85*(3), 275–292.
- Hecht, S. A., Close, L., & Santisi, M. (2003). Sources of individual differences in fraction skills. *Journal of Experimental Child Psychology, 86*(4), 277–302. <https://doi.org/10.1016/j.jecp.2003.08.005>

- Hegarty, M. (2010). Components of Spatial Intelligence. In *Psychology of Learning and Motivation* (Vol. 52, pp. 265–297). Academic Press. [https://doi.org/10.1016/S0079-7421\(10\)52007-3](https://doi.org/10.1016/S0079-7421(10)52007-3)
- Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. *Topics in Cognitive Science*, 3(3), 446–474. <https://doi.org/10.1111/j.1756-8765.2011.01150.x>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- Hofstadter, D. R. (2001). Analogy as the Core of Cognition. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The Analogical Mind: Perspectives from Cognitive Science*. MIT Press. <https://direct.mit.edu/books/edited-volume/2224/chapter/58705/Epilogue-Analogy-as-the-Core-of-Cognition>
- Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 234–259). Oxford University Press.
- Huanca, T. (2008). *Tsimane' Oral Tradition, Landscape, and Identity in Tropical Forest*. Imprenta Wagui.
- Hughes, C., Ensor, R., Wilson, A., & Graham, A. (2009). Tracking executive function across the transition to school: A latent variable approach. *Developmental Neuropsychology*, 35(1), 20–36. <https://doi.org/10/b8bqqv>
- Huizinga, M., Dolan, C. V., & van der Molen, M. W. (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia*, 44(11), 2017–2036. <https://doi.org/10.1016/j.neuropsychologia.2006.01.010>
- Hummel, J. E., & Holyoak, K. J. (2005). Relational Reasoning in a Neurally Plausible Cognitive Architecture: An Overview of the LISA Project. *Current Directions in Psychological Science*, 14(3), 153–157. <https://doi.org/10.1111/j.0963-7214.2005.00350.x>
- Iannone, R., Cheng, J., Schloerke, B., Hughes, E., Lauer, A., & Seo, J. (2023). *gt: Easily Create Presentation-Ready Display Tables*. <https://CRAN.R-project.org/package=gt>
- Ingold, T. (2007). *Lines: A brief history*.
- Ischebeck, A., Schocke, M., & Delazer, M. (2009). The processing and representation of fractions within the brain: An fMRI investigation. *Neuroimage*, 47(1), 403–413. <https://doi.org/10.1016/j.neuroimage.2009.03.041>
- Ishikawa, T., & Newcombe, N. S. (2021). Why spatial is special in education, learning, and everyday activities. *Cognitive Research: Principles and Implications*, 6, 20. <https://doi.org/10.1186/s41235-021-00274-5>
- Jablansky, S., Alexander, P. A., Dumas, D., & Compton, V. (2016). Developmental differences in relational reasoning among primary and secondary school students. *Journal of Educational Psychology*, 108(4), 592–608. <https://doi.org/10.1037/edu0000070>
- Jablansky, S., Alexander, P. A., Eilam, B., Aharon, I., & Sun, Y. (2017). *Test of Relational Reasoning-Junior (TORRjr): Measuring relational reasoning in children and adolescents*.
- Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43), 18243–18250. <https://doi.org/10.1073/pnas.1012933107>
- Jordan, N. C., Hansen, N., Fuchs, L. S., Siegler, R. S., Gersten, R., & Micklos, D. (2013). Developmental predictors of fraction concepts and procedures. *Journal of Experimental Child Psychology*, 116(1), 45–58. <https://doi.org/10.1016/j.jecp.2013.02.001>

- Judd, C. (1908). The relation of special training to general intelligence. *Educational Review*, 36, 28–42.
- Kainulainen, M., McMullen, J., & Lehtinen, E. (2017). Early developmental trajectories toward concepts of rational numbers. *Cognition and Instruction*, 35(1), 4–19.
<https://doi.org/10.1080/07370008.2016.1251287>
- Kalra, P. B., Hubbard, E. M., & Matthews, P. G. (2020). Taking the relational structure of fractions seriously: Relational reasoning predicts fraction knowledge in elementary school children. *Contemporary Educational Psychology*, 62, 101896.
<https://doi.org/10/ghz4t6>
- Kaminski, J. A., & Sloutsky, V. M. (2013). Extraneous perceptual information interferes with children's acquisition of mathematical knowledge. *Journal of Educational Psychology*, 105(2), 351–363. <https://doi.org/10.1037/a0031040>
- Katz, B., Shah, P., & Meyer, D. E. (2018). How to play 20 questions with nature and lose: Reflections on 100 years of brain-training research. *Proceedings of the National Academy of Sciences*, 115(40), 9897–9904. <https://doi.org/10.1073/pnas.1617102114>
- Kidd, J. K., Pasnak, R., Gadzichowski, M., Ferral-Like, M., & Gallington, D. (2008). Enhancing early numeracy by promoting the abstract thought involved in the oddity principle, seriation, and conservation. *Journal of Advance Academics*, 19(2), 164–200.
<https://doi.org/10.4219/jaa-2008-780>
- Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence*, 73, 31–68.
- Kirsh, D. (2010). Thinking with external representations. *AI & SOCIETY*, 25(4), 441–454.
<https://doi.org/10.1007/s00146-010-0272-8>
- Klahr, D., Jirout, J., & Matlen, B. (2013). Children as Scientific Thinkers. In G. F. PhD, G. J. Feist, & M. E. Gorman (Eds.), *Handbook of the Psychology of Science* (pp. 243–247). Springer Publishing Company.
- Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: A training approach. *Review of Educational Research*, 78(1), 85–123. <https://doi.org/10.3102/0034654307313402>
- Klauer, K. J., Willmes, K., & Phye, G. D. (2002). Inducing inductive reasoning: Does it transfer to fluid intelligence? *Contemporary Educational Psychology*, 27(1), 1–25.
<https://doi.org/10.1006/ceps.2001.1079>
- Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, 3(3), 185–225. <https://doi.org/10.1002/acp.2350030302>
- Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., Miller, B. L., & Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia*, 46(7), 2020–2032.
<https://doi.org/10.1016/j.neuropsychologia.2008.02.001>
- Kuhn, D. (2010). What is Scientific Thinking and How Does it Develop? In U. Goswami (Ed.), *The Wiley-Blackwell Handbook of Childhood Cognitive Development* (1st ed., pp. 497–523). Wiley. <https://doi.org/10.1002/9781444325485.ch19>
- Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-Domain Development of Scientific Reasoning. *Cognition and Instruction*, 9(4), 285–327.
https://doi.org/10.1207/s1532690xci0904_1

- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.
<https://doi.org/10.18637/jss.v082.i13>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2020). *lmerTest: Tests in Linear Mixed Effects Models*. <https://github.com/runehaubo/lmerTestR>
- Larkin, J. H., & Simon, H. A. (1987). Why a Diagram is (Sometimes) Worth Ten Thousand Words. *Cognitive Science*, *11*(1), 65–100. [https://doi.org/10.1016/S0364-0213\(87\)80026-5](https://doi.org/10.1016/S0364-0213(87)80026-5)
- Laski, E. V., & Siegler, R. S. (2007). Is 27 a big number? Correlational and causal connections among numerical categorization, number line estimation. *And Numerical Magnitude Comparison Child Development*, *78*(6), 1723–1743.
- Lawson, G. M., & Farah, M. J. (2015). Executive function as a mediator between SES and academic achievement throughout childhood. *International Journal of Behavioral Development*. <https://doi.org/10.1177/0165025415603489>
- Lawson, G. M., & Farah, M. J. (2017). Executive function as a mediator between SES and academic achievement throughout childhood. *International Journal of Behavioral Development*, *41*(1), 94–104. <https://doi.org/10.1177/0165025415603489>
- Le Guen, O., & Balam, L. I. P. (2012). No metaphorical timeline in gesture and cognition among Yucatec Mayas. *Frontiers in Psychology*, *3*. <https://doi.org/10.3389/fpsyg.2012.00271>
- Lee, K., & Bull, R. (2016). Developmental changes in working memory, updating, and math achievement. *Journal of Educational Psychology*, *108*(6), 869–882.
<https://doi.org/10.1037/edu0000090>
- Lee, K., Bull, R., & Ho, R. M. H. (2013). Developmental changes in executive functioning. *Child Development*, *84*(6), 1933–1953. <https://doi.org/10.1111/cdev.12096>
- Lee, K., & Lee, H. W. (2019). Inhibition and mathematical performance: Poorly correlated, poorly measured, or poorly matched? *Child Development Perspectives*, *13*(1), 28–33.
<https://doi.org/10.1111/cdep.12304>
- Lee, V., & Wilkerson, M. (2018). *Data Use by Middle and Secondary Students in the Digital Age: A Status Report and Future Prospects* (pp. 1–43) [Commissioned Paper]. National Academies of Sciences, Engineering, and Medicine, Board on Science Education, Committee on Science Investigations and Engineering Design for Grades 6-12.
https://digitalcommons.usu.edu/itls_facpub/634
- Lehman, D. R., & Nisbett, R. E. (1990). A Longitudinal Study of the Effects of Undergraduate Training on Reasoning. *Developmental Psychology*, *26*(6), 952–960.
<https://doi.org/10.1037/0012-1649.26.6.952>
- Lehto, J. E., Juujärvi, P., Kooistra, L., & Pulkkinen, L. (2003). Dimensions of executive functioning: Evidence from children. *British Journal of Developmental Psychology*, *21*(1), 59–80.
<https://doi.org/10.1348/026151003321164627>
- Leinhardt, G., Zaslavsky, O., & Stein, M. K. (1990). Functions, Graphs, and Graphing: Tasks, Learning, and Teaching. *Review of Educational Research*, *60*(1), 1–64.
<https://doi.org/10.3102/00346543060001001>
- Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means*.
<https://github.com/rvlenth/emmeans>

- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766. <https://doi.org/10/f42jwd>
- Lindenberger, U., Wenger, E., & Lövdén, M. (2017). Towards a stronger science of human plasticity. *Nature Reviews Neuroscience, 18*(5), 261–262. <https://doi.org/10.1038/nrn.2017.44>
- Lloyd, H., Huey, H., Brockbank, E., Padilla, L., & Fan, J. E. (2023). What is graph comprehension and how do you measure it? *Proceedings of the Annual Meeting of the Cognitive Science Society, 45*. <https://escholarship.org/uc/item/6wx5v99w>
- Lövdén, M., Bäckman, L., Lindenberger, U., Schaefer, S., & Schmiedek, F. (2010). A theoretical framework for the study of adult cognitive plasticity. *Psychological Bulletin, 136*(4), 659–676. <https://doi.org/10.1037/a0020080>
- Lüdecke, D., Patil, I., Ben-Shachar, M. S., Wiernik, B. M., Waggoner, P., & Makowski, D. (2021). see: An R package for visualizing statistical models. *Journal of Open Source Software, 6*(64), 3393. <https://doi.org/10.21105/joss.03393>
- Lyons, I. M., & Beilock, S. L. (2013). Ordinality and the nature of symbolic numbers. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 33*(43), 17052–17061. <https://doi.org/10.1523/JNEUROSCI.1775-13.2013>
- Mackey, A. P., Miller Singley, A. T., & Bunge, S. A. (2013). Intensive reasoning training alters patterns of brain connectivity at rest. *Journal of Neuroscience, 33*(11), 4796–4803. <https://doi.org/10.1523/JNEUROSCI.4141-12.2013>
- Mackey, A. P., Miller Singley, A. T., Wendelken, C., & Bunge, S. A. (2015). Characterizing behavioral and brain changes associated with practicing reasoning skills. *PLOS ONE, 10*(9). <https://doi.org/10.1371/journal.pone.0137627>
- Mackey, A. P., Whitaker, K. J., & Bunge, S. A. (2012). Experience-dependent plasticity in white matter microstructure: Reasoning training alters structural connectivity. *Frontiers in Neuroanatomy, 6*, 32. <https://doi.org/10.3389/fnana.2012.00032>
- Maltese, A. V., Harsh, J. A., & Svetina, D. (2015). Data Visualization Literacy: Investigating Data Interpretation Along the Novice—Expert Continuum. *Journal of College Science Teaching, 45*(1), 84–90.
- Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science (New York, N.Y.), 283*(5398), 77–80. <https://doi.org/10.1126/science.283.5398.77>
- Martínez, K., Burgaleta, M., Román, F. J., Escorial, S., Shih, P. C., Quiroga, M. Á., & Colom, R. (2011). Can fluid intelligence be reduced to ‘simple’ short-term storage? *Intelligence, 39*(6), 473–480. <https://doi.org/10.1016/j.intell.2011.09.001>
- Matlen, B. J., Gentner, D., & Franconeri, S. L. (2020). Spatial alignment facilitates visual comparison. *Journal of Experimental Psychology: Human Perception and Performance, 46*(5), 443–457. <https://doi.org/10.1037/xhp0000726>
- Matthews, P. G., Lewis, M. R., & Hubbard, E. M. (2016). Individual differences in nonsymbolic ratio processing predict symbolic math performance. *Psychological Science, 27*(2), 191–202. <https://doi.org/10.1177/0956797615617799>
- McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual

- abilities over the life span. *Developmental Psychology*, 38(1), 115–142.
<https://doi.org/10.1037/0012-1649.38.1.115>
- McKenzie, D. L., & Padilla, M. J. (1986). The construction and validation of the test of graphing in science (togs). *Journal of Research in Science Teaching*, 23(7), 571–579.
<https://doi.org/10.1002/tea.3660230702>
- McMullen, J., Laakkonen, E., Hannula-Sormunen, M., & Lehtinen, E. (2015). Modeling the developmental trajectories of rational number concept(s). *Learning and Instruction*, 37, 14–20. <https://doi.org/10.1016/j.learninstruc.2013.12.004>
- McNeil, N. M., & Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development*, 76(4), 883–899. <https://doi.org/10.1111/j.1467-8624.2005.00884.x>
- Mead, L. A., Mayer, A. R., Bobholz, J. A., Woodley, S. J., Cunningham, J. M., Hammeke, T. A., & Rao, S. M. (2002). Neural basis of the Stroop interference task: Response competition or selective attention? *Journal of the International Neuropsychological Society*, 8(6), 735–742. <https://doi.org/10.1017/S1355617702860015>
- Meert, G., Gregoire, J., & Noel, M. P. (2009). Rational numbers: Componential versus holistic representation of fractions in a magnitude comparison task. *Q J Exp Psychol (Colchester)*, 62(8), 1598–1616. <https://doi.org/10.1080/17470210802511162>
- Meert, G., Gregoire, J., & Noel, M. P. (2010). Comparing the magnitude of two fractions with common components: Which representations are used by 10- and 12-year-olds? *Journal of Experimental Child Psychology*, 107(3), 244–259.
<https://doi.org/10.1016/j.jecp.2010.04.008>
- Michal, A. L., & Franconeri, S. L. (2017). Visual routines are associated with specific graph interpretations. *Cognitive Research: Principles and Implications*, 2(1), 20.
<https://doi.org/10.1186/s41235-017-0059-2>
- Michal, A. L., Shah, P., Uttal, D., & Franconeri, S. (2018). Improving Graph Comprehension With A Visuospatial Intervention. In C. Kalish, M. A. Rau, X. (Jerry) Zhu, & T. T. Rogers (Eds.), *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.
- Miller Singley, A. T., & Bunge, S. A. (2014). Neurodevelopment of relational reasoning: Implications for mathematical pedagogy. *Trends in Neuroscience and Education*, 3(2), 33–37. <https://doi.org/10.1016/j.tine.2014.03.001>
- Miller Singley, A. T., & Bunge, S. A. (2018). Eye gaze patterns reveal how we reason about fractions. *Thinking & Reasoning*, 24(4), 445–468.
<https://doi.org/10.1080/13546783.2017.1417909>
- Miller Singley, A. T., Crawford, J. L., & Bunge, S. A. (2020). Eye gaze patterns reflect how young fraction learners approach numerical comparisons. *Journal of Numerical Cognition*, 6(1), 83–107. <https://doi.org/10.5964/jnc.v6i1.119>
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “Frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41(1), 49–100.
<https://doi.org/10.1006/cogp.1999.0734>
- Moon, J. A., Feng, G., Lentini, J., & Zapata-Rivera, D. (2018). *Facilitating Graph Comprehension Using a Cognitive Model of Successful Solution Processes* (RM-18-07; ETS Research Memorandum Series, pp. 1–9).

- Morrison, F. J., Kim, M. H., Connor, C. M., & Grammer, J. K. (2019). The causal impact of schooling on children's development: Lessons for developmental science. *Current Directions in Psychological Science*, *096372141985566*.
<https://doi.org/10.1177/0963721419855661>
- Murphy, P. K., Firetto, C. M., & Greene, J. A. (2017). Enriching Students' Scientific Thinking Through Relational Reasoning: Seeking Evidence in Texts, Tasks, and Talk. *Educational Psychology Review*, *29*(1), 105–117. <https://doi.org/10.1007/s10648-016-9387-x>
- NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. The National Academies Press.
- Ni, Y., & Zhou, Y.-D. (2005). Teaching and Learning Fraction and Rational Numbers: The Origins and Implications of Whole Number Bias. *Educational Psychologist*, *40*(1), 27–52.
https://doi.org/10.1207/s15326985ep4001_3
- Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science (New York, N.Y.)*, *238*(4827), 625–631. <https://doi.org/10.1126/science.3672116>
- Nothelfer, C., & Franconeri, S. (2020). Measures of the Benefit of Direct Encoding of Data Deltas for Data Pair Relation Perception. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 311–320. <https://doi.org/10.1109/TVCG.2019.2934801>
- Núñez, R., Cooperrider, K., Doan, D., & Wassmann, J. (2012). Contours of time: Topographic construals of past, present, and future in the Yupno valley of Papua New Guinea. *Cognition*, *124*(1), 25–35. <https://doi.org/10.1016/j.cognition.2012.03.007>
- Núñez, R., Cooperrider, K., & Wassmann, J. (2012). Number Concepts without Number Lines in an Indigenous Group of Papua New Guinea. *PLOS ONE*, *7*(4), e35662.
<https://doi.org/10.1371/journal.pone.0035662>
- Núñez, R. E., & Sweetser, E. (2006). With the Future Behind Them: Convergent Evidence From Aymara Language and Gesture in the Crosslinguistic Comparison of Spatial Construals of Time. *Cognitive Science*, *30*(3), 401–450.
https://doi.org/10.1207/s15516709cog0000_62
- Obersteiner, A., Dooren, W., Hoof, J., & Verschaffel, L. (2013). The natural number bias and magnitude representation in fraction comparison by expert mathematicians. *Learning and Instruction*, *28*, 64–72. <https://doi.org/10.1016/j.learninstruc.2013.05.003>
- Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, *55*(3), 169–195.
- O'Shaughnessy, D. M., Cruz, T., Mollica, F., Boni, I., Jara-Ettinger, J., Gibson, E., & Piantadosi, S. T. (2023). Diverse mathematical knowledge among indigenous Amazonians. *Proceedings of the National Academy of Sciences*, *120*(35), e2215999120.
<https://doi.org/10.1073/pnas.2215999120>
- Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: A cognitive framework across disciplines. *Cognitive Research: Principles and Implications*, *3*, 29. <https://doi.org/10.1186/s41235-018-0120-9>
- Padilla, M. J., McKenzie, D. L., & Shaw Jr., E. L. (1986). An Examination of the Line Graphing Ability of Students in Grades Seven Through Twelve. *School Science and Mathematics*, *86*(1), 20–26. <https://doi.org/10.1111/j.1949-8594.1986.tb11581.x>

- Park, Y., & Matthews, P. G. (2021). Revisiting and refining relations between nonsymbolic ratio processing and symbolic math achievement. *Journal of Numerical Cognition*, 7(3), 328–350. <https://doi.org/10.5964/jnc.6927>
- Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology*, 108(4), 455–473. <https://doi.org/10.1037/edu0000079>
- Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31(2), 109–130. <https://doi.org/10.1017/S0140525X08003543>
- Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73–126). Lawrence Erlbaum Associates, Inc.
- Pitt, B., Ferrigno, S., Cantlon, J. F., Casasanto, D., Gibson, E., & Piantadosi, S. T. (2021). Spatial concepts of number, size, and time in an indigenous culture. *Science Advances*, 7(33), 4141–4152. https://doi.org/10.1126/SCIADV.ABG4141/SUPPL_FILE/SCIADV.ABG4141_SM.PDF
- Premack, D. (1983). The codes of man and beasts. *Behavioral and Brain Sciences*, 6(1), 125–136. <https://doi.org/10/fp9m57>
- Purpura, D. J., & Ganley, C. M. (2014). Working memory and language: Skill-specific or domain-general relations to mathematics? *Journal of Experimental Child Psychology*, 122(C), 104–121. <https://doi.org/10.1016/j.jecp.2013.12.009>
- R Core Team. (2013). *R: A language and environment for statistical computing*. <https://www.R-project.org>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org>
- R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ramani, G. B., & Siegler, R. S. (2008). Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. *Child Development*, 79(2), 375–394. <https://doi.org/10.1111/j.1467-8624.2007.01131.x>
- Rattermann, M. J., & Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task. *Cognitive Development*, 13(4), 453–478. [https://doi.org/10.1016/S0885-2014\(98\)90003-X](https://doi.org/10.1016/S0885-2014(98)90003-X)
- Raven, J. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology*, 19, 137–150. <https://doi.org/10.1111/j.2044-8341.1941.tb00316.x>
- Raven, J. (2000). The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology*, 41(1), 1–48. <https://doi.org/10.1006/cogp.1999.0735>
- Redick, T. S. (2019). The hype cycle of working memory training. *Current Directions in Psychological Science*, 0963721419848666. <https://doi.org/10.1177/0963721419848668>
- Ren, K. X., & Gunderson, E. A. (2021). The dynamic nature of children's strategy use after receiving accuracy feedback in decimal comparisons. *Journal of Experimental Child Psychology*, 202. <https://doi.org/10.1016/j.jecp.2020.105015>

- Resnick, I., Davatzes, A., Newcombe, N. S., & Shipley, T. F. (2017). Using Relational Reasoning to Learn About Scientific Phenomena at Unfamiliar Scales. *Educational Psychology Review*, 29(1), 11–25. <https://doi.org/10.1007/s10648-016-9371-5>
- Richland, L. E., & Burchinal, M. R. (2013). Early executive function predicts reasoning development. *Psychological Science*, 24(1), 87–92. <https://doi.org/10.1177/0956797612450883>
- Richland, L. E., Holyoak, K. J., & Stigler, J. W. (2004). Analogy use in eighth-grade mathematics classrooms. *Cognition and Instruction*, 22(1), 37–60. https://doi.org/10.1207/s1532690Xci2201_2
- Richland, L. E., & Morrison, R. G. (2010). Is analogical reasoning just another measure of executive functioning? *Frontiers in Human Neuroscience*, 4. <https://doi.org/10/cv7f88>
- Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children’s development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, 94(3), 249–273. <https://doi.org/10.1016/j.jecp.2006.02.002>
- Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking, and education. *WIREs Cognitive Science*, 6(2), 177–192. <https://doi.org/10.1002/wcs.1336>
- Richland, L. E., Stigler, J. W., & Holyoak, K. J. (2012). Teaching the conceptual structure of mathematics. *Educational Psychologist*, 47(3), 189–203. <https://doi.org/10.1080/00461520.2012.667065>
- Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science (New York, N.Y.)*, 316(5828), 1128–1129. <https://doi.org/10.1126/science.1142103>
- Rinne, L. F., Ye, A., & Jordan, N. C. (2017). Development of fraction comparison strategies: A latent transition analysis. *Developmental Psychology*, 53(4), 713–730. <https://doi.org/10.1037/dev0000275>
- Risko, E. F., & Dunn, T. L. (2015). Storing information in-the-world: Metacognition and cognitive offloading in a short-term memory task. *Consciousness and Cognition*, 36, 61–74. <https://doi.org/10.1016/j.concog.2015.05.014>
- Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/J.TICS.2016.07.002>
- Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science*, 29(8), 1358–1369. <https://doi.org/10.1177/0956797618774253>
- Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2011). Modeling a cascade of effects: The role of speed and executive functioning in preterm/full-term differences in academic achievement. *Developmental Science*, 14(5), 1161–1175. <https://doi.org/10.1111/j.1467-7687.2011.01068.x>
- Rosenberg-Lee, M. (2021). Probing the neural basis rational number difficulties: The role of inhibitory control and magnitude processing. In A. Henik & W. Fias (Eds.), *Learning and education in numerical cognition*. Elsevier.
- Rossi, S., Vidal, J., Letang, M., Houdé, O., & Borst, G. (2019). Adolescents and adults need inhibitory control to compare fractions. *Journal of Numerical Cognition*, 5(3), 314–336. <https://doi.org/10.5964/jnc.v5i3.197>

- Salomon, G., & Perkins, D. N. (1987). Transfer of cognitive skills from programming: When and how? *Journal of Educational Computing Research*, 3(2), 149–169.
- Schneider, M., Beeres, K., Coban, L., Merz, S., Susan Schmidt, S., Stricker, J., & Smedt, B. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science*, 20(3), 1–16. <https://doi.org/10.1111/desc.12372>
- Schrank, F. A., McGrew, K. S., Mather, N., Wendling, B. J., & LaForte, E. M. (2014). *Woodcock-Johnson IV tests of cognitive abilities*. Riverside.
- Shah, P., & Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General*, 124(1), 43–61. <https://doi.org/10.1037/0096-3445.124.1.43>
- Shah, P., Freedman, E. G., & Vekiri, I. (2005). The Comprehension of Quantitative Information in Graphical Displays. In P. Shah & A. Miyake (Eds.), *The Cambridge Handbook of Visuospatial Thinking* (pp. 426–476). Cambridge University Press.
- Shah, P., & Hoeffner, J. (2002). Review of Graph Comprehension Research: Implications for Instruction. *Educational Psychology Review*, 14(1), 47–69. <https://doi.org/10.1023/A:1013180410169>
- Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, 124(2), 209–215. <https://doi.org/10.1016/j.cognition.2012.04.005>
- Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., Susperreguy, M. I., & Chen, M. C. (2012). Early predictors of high school mathematics achievement. *Psychological Science*, 23(7), 691–697. <https://doi.org/10.1177/0956797612440101>
- Siegler, R. S., Fazio, L. K., Bailey, D. H., & Zhou, X. (2013). Fractions: The new frontier for theories of numerical development. *Trends in Cognitive Sciences*, 17(1), 13–19. <https://doi.org/10.1016/j.tics.2012.11.004>
- Siegler, R. S., & Pyke, A. A. (2013). Developmental and individual differences in understanding of fractions. *Developmental Psychology*, 49(10), 1994–2004. <https://doi.org/10.1037/a0031200>
- Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, 62(4), 273–296. <https://doi.org/10.1016/j.cogpsych.2011.03.001>
- Signorell, A. (2023). *DescTools: Tools for Descriptive Statistics*. <https://CRAN.R-project.org/package=DescTools>
- Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Harvard University Press.
- Sjoberg, D. D., Larmarange, J., Curry, M., Lavery, J., Whiting, K., & Zabor, E. C. (2023). *gtsummary: Presentation-Ready Data Summary and Analytic Result Tables*. <https://CRAN.R-project.org/package=gtsummary>
- Smedt, B., Noël, M.-P., Gilmore, C., & Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children’s mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education*, 2(2), 48–55. <https://doi.org/10.1016/j.tine.2013.06.001>

- Smith, C. L., Maclin, D., Grosslight, L., & Davis, H. (1997). Teaching for Understanding: A Study of Students' Preinstruction Theories of Matter and a Comparison of the Effectiveness of Two Approaches to Teaching About Matter and Density. *Cognition and Instruction*, 15(3), 317–393. https://doi.org/10.1207/s1532690xci1503_2
- Smith, L. B. (1984). Young children's understanding of attributes and dimensions: A comparison of conceptual and linguistic measures. *Child Development*, 55(2), 363–380. <https://doi.org/10.2307/1129949>
- Son, J. Y., Blake, A. B., Fries, L., & Stigler, J. W. (2021). Modeling First: Applying Learning Science to the Teaching of Introductory Statistics. *Journal of Statistics and Data Science Education*, 29(1), 4–21. <https://doi.org/10.1080/10691898.2020.1844106>
- St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology*, 59(4), 745–759. <https://doi.org/10.1080/17470210500162854>
- Starr, A., Leib, E. R., Younger, J. W., Project iLead Consortium, Uncapher, M. R., & Bunge, S. A. (2023). Relational thinking: An overlooked component of executive functioning. *Developmental Science*, 26(3), e13320. <https://doi.org/10.1111/desc.13320>
- Starr, A., & Srinivasan, M. (2021). The future is in front, to the right, or below: Development of spatial representations of time in three dimensions. *Cognition*, 210. <https://doi.org/10.1016/j.cognition.2021.104603>
- Starr, A., Vendetti, M. S., & Bunge, S. A. (2018). Eye movements provide insight into individual differences in children's analogical reasoning strategies. *Acta Psychologica*, 186, 18–26. <https://doi.org/10.1016/j.actpsy.2018.04.002>
- Sternberg, R. J. (1977). Component Processes in Analogical Reasoning. *Psychological Review*, 84(4), 353–378.
- Stevenson, C. E., & Hickendorff, M. (2018). Learning to solve figural matrix analogies: The paths children take. *Learning and Individual Differences*, 66, 16–28. <https://doi.org/10.1016/j.lindif.2018.04.010>
- Stricker, J., Vogel, S. E., Schoneburg-Lehnert, S., Krohn, T., Dognitz, S., Jud, N., Spirk, M., Windhaber, M. C., Schneider, M., & Grabner, R. H. (2021). Interference between naive and scientific theories occurs in mathematics and is related to mathematical achievement. *Cognition*, 214, 104789. <https://doi.org/10.1016/j.cognition.2021.104789>
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18, 643–662. <https://doi.org/10.1037/h0054651>
- Swan, M., & Phillips, R. (1998). Graph interpretation skills among lower-achieving school leavers. *Research in Education*, 60(1), 10–20. <https://doi.org/10.1177/003452379806000102>
- Szafir, D. A., Haroz, S., Gleicher, M., & Franconeri, S. (2016). Four types of ensemble coding in data visualizations. *Journal of Vision*, 16(5), 11. <https://doi.org/10.1167/16.5.11>
- Tableau & Forrester. (2022). *Building Data Literacy: The Key to Better Decisions, Greater Productivity, and Data-Driven Organizations* [White Paper]. https://www.tableau.com/sites/default/files/2022-03/Forrester_Building_Data_Literacy_Tableau_Mar2022.pdf

- Taub, G. E., Keith, T. Z., Floyd, R. G., & McGrew, K. S. (2008). Effects of general and broad cognitive abilities on mathematics achievement. *School Psychology Quarterly*, 23(2), 187–198. <https://doi.org/10.1037/1045-3830.23.2.187>
- Taylor, H. A., Burte, H., & Renshaw, K. T. (2023). Connecting spatial thinking to STEM learning through visualizations. *Nature Reviews Psychology*, 2(10), Article 10. <https://doi.org/10.1038/s44159-023-00224-6>
- Thibaut, J.-P., & French, R. M. (2016). Analogical reasoning, control and executive functions: A developmental investigation with eye-tracking. *Cognitive Development*, 38, 10–26. <https://doi.org/10.1016/j.cogdev.2015.12.002>
- Thibaut, J.-P., French, R., & Vezneva, M. (2010). The development of analogy making in children: Cognitive load and executive functions. *Journal of Experimental Child Psychology*, 106(1), 1–19. <https://doi.org/10.1016/j.jecp.2010.01.001>
- Thompson, B., van Opheusden, B., Sumers, T., & Griffiths, T. L. (2022). Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science*, 376(6588), 95–98. <https://doi.org/10.1126/science.abn0915>
- Thompson, R. K. R., & Oden, D. L. (2000). Categorical perception and conceptual judgments by nonhuman primates: The paleological monkey and the analogical ape. *Cognitive Science*, 24(3), 363–396. https://doi.org/10.1207/s15516709cog2403_2
- Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. II. *The Estimation of Magnitudes*. *Psychological Review*, 8(4), 384–395. <https://doi.org/10.1037/h0071280>
- Toledo, R. V. F., Abreu-Mendoza, R. A., & Rosenberg-Lee, M. (2022). Brazilian math teacher’s magnitude representation and strategy use in fraction comparison: A mixed methods study. *PsychArXiv Preprints*. <https://doi.org/10.31234/osf.io/yrngz>
- Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Graphics Press.
- Tufte, E. R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press.
- Tversky, B. (2005). Visuospatial Reasoning. In *The Cambridge handbook of thinking and reasoning* (pp. 209–240). Cambridge University Press.
- Tversky, B. (2011). Visualizing thought. *Top. Cogn. Sci.*, 3, 499–535.
- Tversky, B. (2015). The Cognitive Design of Tools of Thought. *Review of Philosophy and Psychology*, 6(1), 99–116. <https://doi.org/10.1007/s13164-014-0214-3>
- Tversky, B., Kugelmass, S., & Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology*, 23(4), 515–557. [https://doi.org/10.1016/0010-0285\(91\)90005-9](https://doi.org/10.1016/0010-0285(91)90005-9)
- Uncapher, M. R. (2018). Design considerations for conducting large-scale learning research using innovative technologies in schools. *Mind, Brain, and Education*, 6(2), 271–278. <https://doi.org/10.1111/mbe.12185>
- U.S. Department of Education. (2022). *NAEP Report Card: 2022 NAEP Mathematics Assessment*. Institute of Education Sciences, National Center for Education Statistics, National Assessment for Educational Progress (NAEP). <https://www.nationsreportcard.gov/highlights/mathematics/2022/>
- Uttal, D. H. (2000). Seeing the big picture: Map use and the development of spatial cognition. *Developmental Science*, 3(3), 247–264. <https://doi.org/10.1111/1467-7687.00119>

- Valleriani, M., Giannini, G., & Giannetto, E. (Eds.). (2023). *Scientific Visual Representations in History*. Springer International Publishing. <https://doi.org/10.1007/978-3-031-11317-8>
- Vamvakoussi, X., & Vosniadou, S. (2004). Understanding the structure of the set of rational numbers: A conceptual change approach. *Learning and Instruction, 14*(5), 453–467. <https://doi.org/10.1016/j.learninstruc.2004.06.013>
- Van Aken, L., Kessels, R. P. C., Wingbermühle, E., Van Der Veld, W. M., & Egger, J. I. M. (2016). Fluid intelligence and executive functioning more alike than different? *Acta Neuropsychiatrica, 28*(1), 31–37. <https://doi.org/10.1017/neu.2015.46>
- Van Dooren, W., & Inglis, M. (2015). Inhibitory control in mathematical thinking, learning and problem solving: A survey. *ZDM, 47*(5), 713–721. <https://doi.org/10.1007/s11858-015-0715-2>
- Van Hoof, J., Ceulemans, E., & Van Dooren, W. (2021). The Role of the Inhibition of Natural Number Based Reasoning and Strategy Switch Cost in a Fraction Comparison Task. *Studia Psychologica, 63*(1), 64–76. <https://doi.org/10.31577/sp.2021.01.814>
- Van Hoof, J., Degrande, T., Ceulemans, E., Verschaffel, L., & Van Dooren, W. (2018). Towards a mathematically more correct understanding of rational numbers: A longitudinal study with upper elementary school learners. *Learning and Individual Differences, 61*, 99–108. <https://doi.org/10.1016/j.lindif.2017.11.010>
- Van Hoof, J., Verschaffel, L., & Van Dooren, W. (2017). Number sense in the transition from natural to rational numbers. *British Journal of Educational Psychology, 87*(1), 43–56. <https://doi.org/10.1111/bjep.12134>
- Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods, 49*(2), 653–673. <https://doi.org/10.3758/s13428-016-0721-5>
- Vendetti, M. S., Matlen, B. J., Richland, L. E., & Bunge, S. A. (2015). Analogical Reasoning in the Classroom: Insights From Cognitive Science. *Mind, Brain, and Education, 9*(2), 100–106. <https://doi.org/10.1111/mbe.12080>
- Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., & Newcombe, N. S. (2014). Finding the missing piece: Blocks, puzzles, and shapes fuel school readiness. *Trends in Neuroscience and Education, 3*(1), 7–13. <https://doi.org/10.1016/j.tine.2014.02.005>
- Vodegel Matzen, L. B. L., Van Der Molen, M. W., & Dudink, A. C. M. (1994). Error analysis of raven test performance. *Personality and Individual Differences, 16*(3), 433–445. [https://doi.org/10.1016/0191-8869\(94\)90070-1](https://doi.org/10.1016/0191-8869(94)90070-1)
- Vosniadou, S. (2014). Examining cognitive development from a conceptual change point of view: The framework theory approach. *European Journal of Developmental Psychology, 11*(6), 645–661. <https://doi.org/10.1080/17405629.2014.921153>
- Wainer, H. (1992). Understanding Graphs and Tables. *Educational Researcher, 21*(1), 14–23. <https://doi.org/10.3102/0013189X021001014>
- Wechsler, D. (1958). *The measurement and appraisal of adult intelligence (4th ed.)*. Williams & Wilkins Co. <http://content.apa.org/books/11167-000>
- Wendelken, C., & Bunge, S. A. (2010). Transitive Inference: Distinct Contributions of Rostrolateral Prefrontal Cortex and the Hippocampus. *Journal of Cognitive Neuroscience, 22*(5), 837–847. <https://doi.org/10.1162/jocn.2009.21226>

- Wendelken, C., Ferrer, E., Ghetti, S., Bailey, S. K., Cutting, L., & Bunge, S. A. (2017). Frontoparietal Structural Connectivity in Childhood Predicts Development of Functional Connectivity and Reasoning Ability: A Large-Scale Longitudinal Investigation. *Journal of Neuroscience*, *37*(35), 8549–8558. <https://doi.org/10.1523/JNEUROSCI.3726-16.2017>
- Wendelken, C., O’Hare, E. D., Whitaker, K. J., Ferrer, E., & Bunge, S. A. (2011). Increased functional selectivity over development in rostral lateral prefrontal cortex. *Journal of Neuroscience*, *31*(47), 17260–17268. <https://doi.org/10.1523/JNEUROSCI.1193-10.2011>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wickham, H. (2023). *tidyverse: Easily Install and Load the Tidyverse*. <https://CRAN.R-project.org/package=tidyverse>
- Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>
- Wilke, C. O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for ggplot2*. <https://wilkelab.org/cowplot/>
- Wilkerson, M. H., & Laina, V. (2017). Youth Reasoning with Interactive Data Visualizations: A Preliminary Study. *Proceedings of the 2017 Conference on Interaction Design and Children*, 411–416. <https://doi.org/10.1145/3078072.3084302>
- Willingham, D. T. (2008). Critical Thinking: Why Is It So Hard to Teach? *Arts Education Policy Review*, *109*(4), 21–32. <https://doi.org/10.3200/AEPR.109.4.21-32>
- Woltz, D. J., & Was, C. A. (2006). Availability of related long-term memory during and after attention focus in working memory. *Memory & Cognition*, *34*(3), 668–684. <https://doi.org/10.3758/BF03193587>
- World Economic Forum. (2018). *The future of jobs report. Centre for the new economy and society*. <https://doi.org/10.1177/1946756712473437>
- Younger, J. W., O’Laughlin, K. D., Anguera, J. A., Bunge, S. A., Ferrer, E. E., Hoeft, F., McCandliss, B. D., Mishra, J., Rosenberg-Lee, M., Gazzaley, A., & Uncapher, M. R. (2022). *Development of executive function in middle childhood: A large-scale, in-school, longitudinal investigation*. PsyArXiv. <https://doi.org/10.31234/osf.io/xf489>
- Younger, J. W., O’Laughlin, K. D., Anguera, J. A., Bunge, S. A., Ferrer, E. E., Hoeft, F., McCandliss, B. D., Mishra, J., Rosenberg-Lee, M., Gazzaley, A., & Uncapher, M. R. (2023). Better together: Novel methods for measuring and modeling development of executive function diversity while accounting for unity. *Frontiers in Human Neuroscience*, *17*, 1195013. <https://doi.org/10.3389/fnhum.2023.1195013>
- Zhang, L., Fang, Q., Gabriel, F. C., & Szucs, D. (2014). The componential processing of fractions in adults and children: Effects of stimuli variability and contextual interference. *Frontiers in Psychology*, *5*, 981. <https://doi.org/10.3389/fpsyg.2014.00981>
- Zhang, Q., Wang, C., Zhao, Q., Yang, L., Buschkuhl, M., & Jaeggi, S. (2019). The malleability of executive function in early childhood: Effects of schooling and targeted training. *Developmental Science*, *22*(2). <https://doi.org/10.1111/desc.12748>

Appendices

Appendix A: Supplemental Materials for Chapter 2

Additional materials for Section 2.2 can be found online:

Leib, E. R., Starr, A., Younger, J. W., Project iLead Consortium, Bunge, S. A., Uncapher, M. R., & Rosenberg-Lee, M. (2023). Testing the whole number interference hypothesis: Contributions of inhibitory control and whole number knowledge to fraction understanding. *Developmental Psychology*, *59*(8), 1407–1425.
<https://doi.org/10.1037/dev0001557>

Supplementary materials: <https://doi.org/10.1037/dev0001557.supp>

Data cleaning and analysis scripts are available on Open Science Framework:
<https://osf.io/z7hxa/>

Appendix B: Supplemental Materials for Chapter 3

Additional details about memory task

Stimuli. The stimuli consisted of four sets of eight AI-generated adapted faces from [Generated.Photos](#), one set for each of the four trials. Faces were chosen such that it was plausible that the individuals could have lived in a nearby Bolivian town. Each set of eight faces had two male children, two female children, two female adults, and two male adults. We selected older- and younger-looking faces within each age+gender subgroup (e.g., older and middle-aged adult woman and older and middle-aged adult man, etc.) to make sure that the two faces within each subgroup did not look too much alike.

Using Adobe Photoshop, the backgrounds of the images were removed to make them more uniform. Additionally, the shirt colors were changed to match one other individual in the set, such that there were only four shirt colors in each set. In addition to making the images simpler, changing the shirt color also added a nonrelevant visual feature that participants could have grouped by that was not inherent to the individuals' faces like age and gender are. Therefore, in addition to sequential order and grouping by preference, participants could have also organized cards in age order or grouped by age, gender, or some combination, or grouped by shirt color, though no participants organized the cards in these alternative ways. Faces were pseudo-randomly assigned to one of the four shirt colors with the constraint that exactly two individuals had each color and those two individuals could not be from the same age+gender subgroup (i.e., both adult men could not have the same color shirt). Shirt colors were different in each set.

The faces were scaled to 2.5 x 2.5 inches, printed on card stock, and laminated. The order of the faces in each trial was randomized before the experiment, and all participants saw the same card order, regardless of condition. Fig. 1A shows the face stimuli for the first trial and Fig. S1 shows all trials.

Condition-Specific Details.

Order Condition. At the beginning of the task, the experimenter explained (in Spanish, subsequently translated into Tsimane'), "We are going to play a memory game with these face cards. These people are going to the market today. I'm going to give you each person, one by one, and tell you the order in which they arrive at the market. I want you to remember this order, because afterwards I'm going to ask you about this order. To help you remember the order, you can put the cards on the table if you want." After the translator restated the instructions in Tsimane', the experimenter started handing the participant cards one by one, saying the order. For example, on the first trial the experimenter said, "This boy arrived first. Next, this man arrived. Next, this girl arrived. After this woman arrived..." and so on (see Fig. 1A and S1 for order).

For the memory questions we wanted to ensure that it was not possible to answer the questions without considering the full set of cards, so we asked conjunctive questions about order and either age group or gender. On the first trial, participants were asked to point to the first child to arrive (Fig. 1A and S1, card 1) and the last adult to arrive (Fig. 1A and S1, card 7). See Table S3 for counterbalancing across trials and the full list of questions and correct answers.

Preference Condition. At the beginning of the Preference condition, the experimenter explained (in Spanish, subsequently translated into Tsimane'), "We are going to play a memory game with these face cards. I'm going to give you each person, one by one, and tell you about their preferences. I want you to remember these preferences, because afterwards I'm going to ask you about them. To help you remember the preferences, you can put the cards on the table if you want to." Then the experimenter said what the two preferences were and started handing the participant cards one by one, saying each card's preference.

For example, on the first trial the experimenter said, "To begin, I'm going to tell you about these people's preferences between eating plantain and eating coconut. This boy prefers plantain. This man prefers coconut. This girl prefers coconut. This woman prefers plantain..." and so on (see Fig. 1A and S1 for preferences). Each trial gave preferences between two options, which were developed in collaboration with the translators to be gender-neutral and age-neutral, such that participants could not use gender or age as deterministic cues. See Fig. S1 for all the preferences. Note that the instructions only said to remember the preferences, never using the term "groups."

For the memory questions, we asked conjunctive questions about preference and either age group or gender. For example, on the first trial participants were asked to point to all the adults who preferred plantains (Fig. 1A and S1, cards 4 and 5) and then all the children who preferred coconut (Fig. 1A and S1, cards 3 and 8). See Table S3 for counterbalancing across trials and the full list of questions and correct answers.

Preference group assignment was pseudo randomized, with two males and two females in each group. In terms of the sequence in which cards were handed out, all four cards in a preference group were never presented in a row, and three cards of the same preference were not presented in a row across two consecutive trials.

Control Condition. At the beginning of the no-task Control condition, the experimenter explained (in Spanish, subsequently translated into Tsimane'), "We are going to do something very simple with these face cards. I am going to give you the cards and you are going to put them on the table in any way that you want." Then the experimenter started handing the participant cards one by one, without saying anything. Note that participants were handed the cards in the same order as in the other two conditions but were not prompted to remember anything about the cards.

For the questions, the participant was simply asked to point to a card of one demographic dimension (age group or gender). For example, on the first trial the participant was asked to point to an adult and then to a child. The demographic dimension was the same as what was asked about in the Order and Preference condition questions for each trial (Table S3).

This condition included only two trials instead of four. Participants with odd participant IDs received trials 1 and 3, and participants with even IDs received trials 2 and 4⁴.

Coding Organization of the Layouts. After shape was coded, the *organization* of each card layout was coded by the first author as either "sequentially ordered," "grouped," or "neither." For a layout to be coded as sequentially ordered, it needed to preserve the sequence that the cards were handed out in, regardless of condition. For example, if the layout was a 2x4 grid and the first four cards were placed in order from right to left, then the second row needed to have the next four cards in order, too, either again from right to left (e.g., Fig. S2E), or alternatively from left to right (e.g., Fig. S2F). In the second case, the sequential order "snakes through" the shape, meaning that one could place their finger on the first card and move along the cards in order through the 8th card without needing to lift their finger. In either case, the path through the cards preserved the sequential order (Cooperrider et al., 2017). See Fig. S2 for examples of various layouts that were coded as sequentially ordered.

⁴ Some early Control participants were given all four trials or only trials 1 and 2 or 3 and 4. However, only the first two trials that they completed were included in analysis.

A layout was coded as grouped if a straight line could be drawn (horizontal, vertical, or diagonal) that perfectly separated the cards belonging to each preference⁵. This was coded independent of whether there was a visible gap between the two groups in the shape. For the Preference condition trials, we also coded for “attempted grouping,” which was when it was clear from the card placement and the participant’s responses to the memory test questions that they had been attempting to use a spatial grouping strategy, but made a small execution error (i.e., misplaced one card, or swapped two cards with each other), resulting in the final layout not being correctly grouped. Even with this additional code, the organization coding was conservative and likely underestimates the prevalence of the spatial grouping strategy because of the strict requirement for the grouping to be correct in order to be coded as grouped or one card away from correct to be coded as attempted.

A few participants placed cards face down on the table, so the backs of the cards were visible in the trial images (3 participants on the first trial, 4 total, accounting for 5 trials across the whole task). Organization was able to be coded for these trials because the numbering system the experimenter used to order the cards between trials was written on the back (though it was not a system recognizable to participants). For six participants (accounting for 13 total trials) who made piles of cards (either 1, 2, or 3 piles) it was not possible to code organization from the images alone. For two of these participants (accounting for 8 trials), we were able to code the organization from video. However, the other four participants, all in the Control condition, did not have videos (3 made piles on their first trial, and 1 on their second trial). These participants were removed from all analyses of organization, since we were not able to code the organization of the card layouts.

Accuracy Scoring. For the response to a memory test question to be scored as correct, the given response had to satisfy both parts of the question. For example, in the first trial of the Order condition, the second question was “Who was the last adult to arrive?” so the correct answer (Fig. 1A, card 7) had to satisfy both “adult” and “last”. In the Preference condition, the first question on the first trial was “Who are all the adults who prefer plantain?” and the correct answer (Fig. 1A, cards 4 and 5) had to satisfy both “adult” and “prefer plantain.” See Table S3 for the memory test questions and correct answers. The accuracy score for each trial was the sum score of the two memory questions, and therefore could be either 0 (both wrong), 1 (one correct), or 2 (both correct).

Note that for each Preference condition question, both cards needed to be given for the responses to be scored as correct, and only those two cards. Sometimes, participants responded with all four cards of the asked-for preference. In the case studies, we refer to these kinds of responses as “partially correct,” but we were conservative in scoring accuracy and still scored these partially correct responses as incorrect (0) because they did not satisfy both parts of the question. Therefore, our accuracy measure likely underestimates memory performance for preferences.

Responses to the Control condition questions were not scored for accuracy since they were not memory questions.

⁵ There was one exception to this rule. On one trial a participant was grouping by preference but was placing the cards on the edge of the table and they started falling on the ground. The translator prompted the participant to move the cards onto the table more. The participant continued grouping but with the rest of the cards for that group in a different location on the table. In the resulting card layout, a straight line cannot be drawn to separate the groups, but the video evidence is clear that the participant was correctly using a spatial grouping strategy during the whole trial and correctly grouped by preference, so the organization for this trial was coded as grouped.

Missing Data, Data Cleaning, and Handling Experimenter Errors. For two participants—one child and one adult, both in the Order condition—there was no photo of their card layouts on the fourth trial. The child participant’s session was video recorded, so we used a screenshot of the video to code the organization and shape for that trial. However, the adult participant did not have a recorded session. Therefore, the organization and shape could not be coded. This missing photo did not affect any of the analyses that included only the first trial, but for the analyses that included all trials, this participant was removed from analysis. One adult woman in the Order condition was distracted (nursing her baby) during the second trial of the task, so the experimenter decided to terminate the task after the second trial. Data for her first trial was included in analyses, but data from the second trial was removed, and she has no data for trials three and four. Therefore, there were two adult participants in the Order condition who were included in analyses of T1, but excluded for analyses of all trials, resulting in a sample size of 12 children and 15 adults (27 total) for these analyses.

Two adults in the Control condition were not asked the questions following one or both of their trials. However, responses to the Control condition questions were not analyzed so these missing data had no impact on analysis. Finally, as described above, three Control condition participants were removed from analyses of organization on T1 (2 children and 1 adult) because the organization of the layouts could not be coded from the trial images, resulting in a Control condition sample size of 19 children and 28 adults (47 total) for these analyses.

There were a few trials on which the experimenter made an error in ordering the face stimuli while setting up the task, and therefore handed the cards to the participant in the incorrect order (3 participants, accounting for 5 trials total, all Order condition participants). Specifically, the cards were distributed in the order 12745638, with cards 3 and 7 switched. If a participant placed the cards down in that order, then the organization was coded as sequential since it was the sequence given to the participant. There was also one trial in the Preference condition during which the experimenter accidentally switched the preferences of the first two cards, though the participant did not group by preference.

Additional details about demographic intake survey

Age. Participants were asked their age and birthdate. When both pieces of information were given, age at the time of testing was calculated from the birthdate and used in analyses. If a participant reported their age but did not know their date or year of birth (22 participants), their reported age was used in analyses. Three of these participants said they did not know their age or birthdate; for these participants, the research coordinator estimated their age.

Schooling. Participants were asked to report the number of years of schooling they had completed. There were two participants who said they did not know (one in the Preference condition and one in the Order condition), so we recorded NA for their years of school. These participants were removed from the supplemental analyses involving schooling as a predictor. Note that there is a great deal of variability within this self-reported measure because what constitutes a “completed year of school” may vary widely across participants. For example, some participants may have attended school only once per week for 2 hours and reported that as one year of school. Further, what schooling entails differs depending on the individual’s age and what community they are from. This measure also relies on the individual knowing how to count, which is not the case for all our participants. However, there are no written records of attendance, so self-report is our best approximation of years of schooling.

Data Analysis

Fisher's Exact Test was used for all cases testing the association between two categorical variables, including follow-up pairwise comparisons between conditions. Unlike χ^2 Test and other traditional hypothesis tests, Fisher's Exact Test does not calculate a test statistic or use a sampling distribution. Rather, Fisher's Exact Test calculates all possible contingency tables that have the same marginal distribution as the observed table, and then calculates an exact p-value by taking the proportion of possible tables that are as extreme or more extreme than the observed table. Therefore, the only output from the test is a p-value. When Fisher's Exact Test is run on a 2 x 2 contingency table, an effect size—an odds ratio (OR)—and its 95% confidence interval can also be calculated from the contingency table. The OR gives the odds of an outcome occurring in one condition (e.g., sequentially ordering in the Order condition) relative to the odds of that outcome occurring in another condition (e.g., sequentially ordering in the Control condition). When the odds are similar, the OR will be close to 1, but when the odds are different, the OR will be much larger than 1 or much smaller than 1, though cannot be smaller than 0. Note that when one of the cells of observed data contains a 0, the OR will either be 0 if that cell is in the numerator of the ratio or infinity if it is in the denominator.

All analyses were run in R version 4.2.1 (R Core Team, 2022). We ran Fisher's Exact Test using `fisher.test` from `stats` 4.2.1 (R Core Team, 2022) and report the p-value. When Fisher's Exact Test was used on a 2x2 contingency table, we also report the OR and report the 95% confidence interval in the supplement. When the contingency table was too large for this exact calculation, the p-value was calculating by Monte Carlo simulation using 10000 replicates, which we specified as arguments in the `fisher.test` function. In follow-up pairwise comparisons, we used Bonferroni correction to correct for multiple comparisons, multiplying the p-value by 3 since we did three pair-wise tests (Order vs Control, Order vs Preference, and Preference vs Control). For cases in which a χ^2 Test of Independence was appropriate (i.e., when all cells had expected values greater than 5), we also report the results of the χ^2 test in the Supplementary Text. All χ^2 tests showed the same pattern of results as the Fisher's Exact Tests.

T-tests were run using the `t.test` function from `stats` 4.2.1 (R Core Team, 2022). Logistic mixed effects models were run using the `glmer` function from `lme4` 1.1.33 (Bates et al., 2023) and `lmerTest` 3.1.3 (Kuznetsova et al., 2020). Logistic regressions were run using the `glm` function from `stats` 4.2.1 (R Core Team, 2022). `Tidyverse` 2.0.0 (Wickham, 2023) packages were used to wrangle and clean the data. `ggplot2` 3.4.2 (Wickham, 2016), `cowplot` 1.1.1 (Wilke, 2020), and `ggh4x` 0.2.6 (Brand, 2023) were used to visualize the data. `DescTools` 0.99.50 (Signorell, 2023) was used to calculate the multinomial confidence intervals for Fig. 3A. `gt` 0.10.0 (Iannone et al., 2023) and `gtsummary` 1.7.2 (Sjoberg et al., 2023) were used to make the supplementary tables. `papaja` 0.1.2 (Aust & Barth, 2022) and `english` 1.2.6 (J. Fox et al., 2021) were used to format and output the results from R. Data and analysis files are available on Open Science Framework at <https://osf.io/75vrj/>.

Additional details about results

Participants spatially organized cards to represent relevant information on the first trial.

Post hoc tests reveal pairwise differences in organization between conditions. For both children and adults, we observed a significant effect of condition on organization (Fisher's Exact Test: $ps < .001$). This significant result warranted post hoc follow-up tests to examine how the distribution of organizations differed between conditions. First, we tested whether participants sequentially ordered more in the Order condition than in each of the other two conditions (two pairwise comparisons per age group). Second, we tested whether participants grouped by preference more in the Preference condition than by chance in the Control condition: that is, whether they placed cards 1-4-5-6 together and 2-3-7-8 together without any preference information (one pairwise comparison per age group). In these analyses, we collapsed organization into a binary variable—either sequential and not-sequential or grouped and not-grouped--

depending on the question. This made the analyses 2 (conditions) x 2 (organizations), and allowed the odds ratios (i.e., the effect size) to be calculated and reported. Given that we used three pairwise comparisons, we used the Bonferroni method to correct for multiple comparisons, multiplying the outputted p-value from the analysis by 3, the number of comparisons. These corrected p-values are reported in this section (BC-*p*).

In the Order condition, 11/12 children and 13/17 adults represented the sequential order in their layouts, significantly more than children and adults in the Control condition (children: 5/19; adults: 9/28; Table S4) and in the Preference condition (children: 4/13; adults: 4/15; Table S4).

In the Preference condition, 9/15 adults spatially grouped the cards by preference—significantly more than in the Control condition, in which none of the 28 adults organized cards in this way ($OR = \infty$, 95% CI [6.48, ∞], BC-*p* < .001). Only 2/13 children grouped in the Preference condition, and none of the 19 children grouped in the Control condition. Although these proportions are low and are not statistically different ($OR = \infty$, 95% CI [0.28, ∞], BC-*p* = .472), based on various sources of evidence we believe that the two Preference condition children who grouped did so intentionally and therefore were behaving differently than the children in the Control condition. First, the shape that one child used to represent the grouped organization was two spatially separated clusters of cards. We interpret adding space between the preference groups as demonstrating intentional grouping. The other child used a line shape, so grouping was not apparent from visual inspection alone, but was evident in the way he used his card layout to answer the memory questions. He responded to the test question “all the adults who prefer plantains” by pointing to the four cards on the left, which corresponded to all four individuals who preferred plantains, and answered “all the children that prefer coconut” with the four cards on the right, which corresponded to all four individuals who preferred coconut. Therefore, even though he did not answer the questions fully correctly, he correctly identified which faces had each preference. This partially correct and spatialized response behavior suggests that he had intentionally represented the preference groups. Finally, it is statistically unlikely that cards randomly placed would result in the correctly grouped organization. For example, in the Control condition, in which participants had no knowledge of preferences, cards with the same preferences were grouped together in only 2 out of the 95 layouts⁶ (2.11%) over the two trials. Together, it seems most likely that these two children were behaving differently than children in the Control condition, even though we do not have the statistical power to detect this difference with the small sample size. Although a logistic regression showed that adults grouped more than children ($b = 0.45$, 95% CI [0.11, 0.78], $t = 2.61$, $p = .015$; Fig. 3A), we did not find an effect of age on grouping when age was treated as a continuous variable (see section on testing for effects of age, schooling, and literacy).

Relating organization to memory accuracy. We examined whether participants who sequentially ordered in the Order condition or grouped in the Preference condition were more accurate on the memory questions than participants in those conditions who did not use these organizational strategies. We first analyzed only T1 because on this trial participants had not yet heard any questions, meaning that when they placed the cards they did not know what they would be asked for or the structure of the questions. Then, we analyzed all trials together to confirm that the pattern found for T1 held for all trials. For each condition, we used a logistic mixed effects model predicting accuracy (correct or incorrect) on each memory question (two per trial) from the question number, trial number, and organization, with a random effect for participant. Only participants with data for all trials were included in these analyses (see *SI*

⁶ There were 50 total participants in the Control condition, yielding 100 layouts over the two trials. Of those 100 layouts, 5 could not be coded for organization and we do not have video to verify the organization. See Methods.

Methods section on missing data handling). We chose to analyze the two conditions separately because of near-ceiling effects in the Order condition, in which most participants sequentially ordering the cards.

On T1 of the Preference condition, participants who grouped by preference were more accurate on the test questions than those who did not group ($\Delta M = 1.03$, 95% CI [0.47, 1.60], $t(11.11) = 4.01$, $p = .002$). Across all trials, this pattern held: the odds of answering a memory question correctly were 31.54 times greater when the card organization was grouped versus not grouped ($b = 3.45$, 95% CI [2.21, 4.69], $z = 5.47$, $p < .001$; Table S5). We interpret these results as additional evidence that participants were using space strategically, since grouping predicted better accuracy. It is worth noting that because we were conservative in our accuracy scoring, we are likely underestimating the effect of organization. For example, if a participant responded with all cards of a given preference, the response was marked as incorrect. Therefore, the accuracy score underestimates participants who answered in this way. It also underestimates participants who attempted to group, but made an error, because when they answered based on their layout the response was incorrect. Thus, finding that organization predicts accuracy even using a strict definition of correct, suggests that the effect may be even stronger. Future studies should examine this relation with a larger sample size and more sensitive measures of accuracy.

The Order condition analyses are more difficult to interpret because of the near-ceiling effect of participants sequentially ordering cards (24/29). While we ran the analyses for completeness, we use caution in interpreting the results. There was no difference in accuracy between participants who sequentially ordered and those who did not, neither on T1 ($\Delta M = 0.01$, 95% CI [-0.53, 0.55], $t(5.53) = 0.04$, $p = .971$) nor across trials ($b = -0.13$, 95% CI [-1.18, 0.92], $z = -0.24$, $p = .810$; Table S6). The mixed effects model shows that participants were more accurate on questions about “first” than questions about “last” (Table S6).

Testing for effects of age, schooling, and literacy on card organization. Given that many participants did have some schooling, it is central to the interpretation of the results to confirm that strategically organizing the cards on T1 (i.e., sequentially ordering in the Order condition and grouping in the Preference condition) was not simply due to experience in school or being literate. We used logistic regression to predict organization from schooling and literacy. We also controlled for age, adding it as a continuous predictor, and therefore collapsed across age groups for these analyses. We chose to analyze the two conditions separately because of the large proportion of participants in the Order condition who sequentially ordered the cards, and because there may be different effects of schooling and literacy on performance on each condition.

In the Preference condition, neither years of schooling ($b = 0.07$, 95% CI [-0.29, 0.53], $z = 0.35$, $p = .727$) nor literacy ($b = 0.21$, 95% CI [-0.04, 0.49], $z = 1.59$, $p = .112$) predicted grouping the cards by preference. Further, although in Fig. 3A it may look like there is a difference in grouping between children and adults, we did not find an effect of age when using age as a continuous variable. The two children who grouped were two of the oldest children (ages 13 and 14); however, the other three children in this older age range (13-14) did not group. Future work should test for possible maturational or cohort effects with a larger sample. Given this potential difference between children and adults, we look more qualitatively at the relation between schooling, literacy, and organization in this condition. On average, the children in this condition had slightly more schooling than the adults ($\Delta M = 2.07$, 95% CI [0.06, 4.07], $t(22.70) = 2.13$, $p = .044$), even though they were younger. Further, children were presently attending school, whereas adults were no longer attending school. Therefore, if schooling were a determining factor for strategically using space, we should have observed more children grouping than adults, but this is not what we found. Instead, among the participants who grouped on T1 we observed a range of schooling, including an adult

who reported 0 years of school, and three adults who reported just 3 years of school. For literacy, there was no difference between children and adults in this condition ($\Delta M = -1.12$, 95% CI [-3.86, 1.62], $t(25.97) = -0.84$, $p = .409$). The set of participants who grouped on T1 (11 total) was made up of four individuals who were not literate and seven individuals who were.

Due to the near-ceiling effect of sequentially ordering the cards in the Order condition (24/29), there was not enough variation in organization to test for the effects of age, schooling, and literacy. It is worth noting that all five participants who did not sequentially order were also not literate. However, literacy was not fully determinant because there were nine additional non-literate participants in this condition who did make sequential orders on T1. For example, there were four participants in this condition who reported 0 schooling and were not literate (see Fig. S5 and Section 2.5). Of these four, two sequentially organized the cards on T1 and the other two did not. Dovetailing with the results of the Preference condition analysis, the set of participants who sequentially ordered on T1 was made up of a set of individuals with varied characteristics: some that were literate, others that were not, some that had more schooling, and some that had none. Together, these results suggest that although it is possible, even likely, that schooling and literacy have some influence on task performance, these variables do not fully determine whether participants spontaneously offloaded task-relevant information to space.

Shape of card layouts varied within and between conditions. A Fisher's Exact Test on a 3 (condition) x 9 (number of shapes) contingency table confirmed that the distribution of shapes differed between conditions ($p = .006$). Some shapes were found in some conditions but not others. For example, square (Fig. 2C) was used by eight participants in the Control condition (16.0%), but not by any participants in the Order or Preference conditions (Fig. S3). Another example is two clusters (two sets that were separated by a gap in space; Fig. 2H, I): 4 of the 11 participants (36.4%) who grouped in the Preference condition placed cards in this shape, but no participants in the Order condition and only one in the control condition placed cards in this way. Note that shape was coded blind to condition and organization—coders were simply judging whether there was enough space between two sets or lines to visually look like two separate clusters. By contrast, lines and rectangles were used reliably across all three conditions.

Specifically, lines were the most common shape created in each condition. We had not expected to find that lines would be used to organize by preference. It is interesting to note that although using this shape offloads part of the memory demand—cards with like-preferences are closer to each other than those with different preferences—participants still had to remember the group boundary because it was not visually apparent (i.e., there was no visible gap between the groups).

Given the prevalence of this shape, we compared the proportion of participants on T1 who created lines versus all other shapes between conditions. For adults, the distribution of lines versus all other shapes differed between conditions (Fisher's Exact Test on a 3 (conditions) x 2 (line vs not line) contingency table: $p < .001$). Post hoc Fisher's Exact Tests on 2 (pairwise conditions) x 2 (line vs not line) contingency tables revealed that this effect was driven by the difference between the Order and Control conditions. The odds of an adult in the Order condition creating a line were 14.63 times higher than the odds of creating a line in the Control condition (95% CI [2.79, 108.35], $BC-p < .001$). The proportion of lines used in the Preference condition was between the Order and Control condition proportions and did not significantly differ from either ($OR = 3.51$, 95% CI [0.69, 20.43], $BC-p = .467$ and $OR = 4.17$, 95% CI [0.67, 32.07], $BC-p = .296$, respectively). Therefore, we cannot conclude whether participants used lines more in the general context of a memory demand or specifically in response to an ordering prompt. That said, the odds of a line being sequentially ordered were 14.00 times higher in the Order condition than in the Preference condition (95%

CI [0.85, 972.43], $p = .036$). Together, these results suggest that adults did not default to using lines, and specifically used sequentially ordered lines in the Order condition.

In contrast with adults, children created lines equally often across all three conditions (Fisher's Exact Test on a 3 (conditions) \times 2 (line vs not line) contingency table: $p = .599$). About half of all shapes created in each condition were lines, suggesting that many children may default to creating lines regardless of the prompt. One possible explanation for this observation is that it may be due to their current ongoing exposure to schooling. However, as noted above, the creation of a line did not mean that it was sequentially ordered. This was the case in the Control condition, where only one third of lines were sequentially ordered, whereas all lines created in the Order condition were sequentially ordered ($OR = \infty$, 95% CI [1.09, ∞], $p = .029$). We interpret these results across age groups as indicating that children and adults were sensitive to the memory prompt when creating lines, selectively using sequentially ordered lines in the Order condition.

Variability in card direction among sequentially ordered layouts. In addition to coding the organization and shape, we also coded the directionality that the cards were placed in, relative to the participant. Directionality is most interesting when participants preserved the sequential order of the cards, regardless of the condition. For example, if a participant placed cards in a line, starting from her left side and moving towards her right, the direction was coded as "left-to-right" (e.g., Fig. S2A). Out of the 46 participants across conditions who captured sequential order on T1, 23 created lines, and 20 of the lines were left-to-right (Fig. S3A), compared to three right-to-left lines (Fig. S2B). This bias towards left-to-right lines was surprising given prior work with the Tsimane' that found no preference between right-to-left and left-to-right linear orderings (Pitt et al., 2021). Further, no participants made vertical lines (e.g., top-to-bottom) for sequential order.

Aside from lines, 19 other participants captured sequential order with two or more rows and columns (i.e., line + extra, rectangle, square, and snake shapes; Fig. S2C-J, L). Here, there was greater variability in card directionality. For example, two participants first placed a row of cards left-to-right, then below that, placed a second row left-to-right (e.g., Fig. S2D). Three participants placed the first row left-to-right, and then placed the second row above the first left-to-right (e.g., Fig. S2I). One participant placed cards right-to-left, with the second row above the first (Fig. S2E), and still another participant placed a column of cards top-to-bottom, then to the left of it placed another top-to-bottom column (Fig. S2H). Other participants had their sequential order "snake through" the shape, such that one could place their finger on the first card and not pick it up until reaching the last card (Fig. S2C, F, G, and J). This variability demonstrates the spatial flexibility that participants had while working with the cards and further supports the conclusion that participants were innovating ad hoc strategies on this task.

Use of spatial strategies increased after the first trial. In addition to testing for the spontaneous use of space on T1, we also examined strategy change across the four trials. In the Order condition, most participants sequentially ordered on T1, leaving little room for strategy change over the subsequent trials (Fig. S4). Across children and adults, only one participant in this condition, an adult, did not sequentially order on any trial. This means that all children in this condition sequentially ordered the cards on at least one trial. Although there was one child on each trial that did not sequentially order, it was not the same child each time. A mixed effects logistic regression predicting sequential organization (1: sequentially ordered, 0: not sequentially ordered) from a fixed effect of trial and random effect of participant confirmed that there was no change in sequential ordering over the trials for children ($b = 0.00$, 95% CI [-1.29, 1.29], $z = 0.00$, $p > .999$). For adults, the results revealed a small effect of trial on the odds of sequentially ordering ($b = 16.87$, 95% CI [0.62, 33.12], $z = 2.04$, $p = .042$).

In the Preference condition, in addition to correctly grouping, some participants attempted to group; that is, it was clear from their card placing behavior that they were intending to organize the cards by preference but ultimately made an error when placing one or two cards, resulting in the final layout being incorrectly grouped (e.g., Fig 5, top row, trial 3). Critically, all participants who had a layout that was coded “attempted group” had either already correctly grouped on a previous trial, or correctly grouped on the subsequent trial, further suggesting that this was intentional use of a spatial grouping strategy. Therefore, for this analysis we combined grouped and attempted group because the same spatial strategy was being used even when it was not executed perfectly. To test for strategy change, we used a mixed effects logistic regression predicting grouped organization (1: correctly grouped or attempted group; 0: not grouped) from a fixed effect of trial and random effect of participant. For children, this analysis confirmed that the odds of using a grouped organization increased after T1 ($b = 1.71$, 95% CI [0.19, 3.23], $z = 2.21$, $p = .027$; Fig. 4). For adults, the effect of trial did not reach significance ($b = 0.83$, 95% CI [-0.03, 1.68], $z = 1.90$, $p = .058$), however, this model explained significantly more variance than the null model ($\chi^2(1) = 4.56$, $p = .033$). This in part may be because more than half of the adults already grouped on T1, leaving less room for improvement than children had. In addition to the increased use of the grouping spatial strategy, accuracy on the memory questions also increased over trials (Table S5). The results from the Preference condition provide evidence of strategy change.

Detailed descriptions of Preference condition case studies. MP was approximately 40 years of age. On the first trial, she grouped the cards by preference and answered both memory questions correctly. The shape created by the cards was coded as rectangle by both coders, indicating that they perceived the cards as one cluster (i.e., no spatial separation between the two groups). On the second trial, she again grouped the cards by preference and answered both memory questions correctly. However, on this trial, the shape the cards made was coded as two clusters by both coders, indicating they perceived two separate clusters of cards with a gap in space between them (Fig. 5, top row, trial 2). It is important to note, as described in the Methods, that the two coders were blind to condition and coded layouts in a random order; therefore, they were not influenced by expectations of what information the participant may have been trying to represent nor by that participant’s other trials. On the third trial, MP made an error while placing the cards by preference, so the final grouping was not correct, but it was clear that she was attempting to group based on the pattern of her card placing and the resulting two-cluster shape (Movie S2). On the fourth trial, she correctly grouped and again answered the questions correctly, again creating two clusters of cards to represent the preference types. Further, in this last trial, the groups are further apart than in the previous trials. We interpret this trajectory of performance, specifically the increase in spatial separation between groups, as suggesting that MP was improving or fine-tuning her strategy—even though she was already performing accurately—possibly to complete the task more efficiently by offloading even more of the memory demand.

AS was 58 years of age. Unlike MP, AS did not spontaneously group cards by preference on the first trial. She started with a common layout (Fig. 5, middle row, trial 1: 87123456), which was neither sequentially ordered nor grouped. On the second trial, she sequentially ordered the cards. She did not answer any of the memory questions correctly on the first or second trial. On the third trial, she attempted to group by preference, but made an error when placing the cards, resulting in an organization that was not correctly grouped. For this trial, the prompt was to remember people’s preferences between traveling in a canoe versus callapo (another type of boat). After placing the first two cards in the center from left to right (canoe, then callapo), AS placed the third card (callapo) on the left, next to the canoe-preferring first card. However, after this third card, the rest of the cards were sorted correctly, with all remaining callapo-preferring cards on the left and canoe-preferring cards to the right (Fig. 5, middle row, trial 3). For the memory questions, AS spatialized her responses, selecting the five cards on the right to answer “all the

children that preferred canoe” and the remaining three cards on the left to answer “all the adults that preferred callapo.” Even though these responses were incorrect (both because the cards were not correctly grouped and because she did not limit her responses to only children or adults), it was clear from this response pattern that she was now intending to group by preference on this trial. Interestingly, AS seemed to struggle with the boundary between the groups: when answering the first question, she started from the right, moving left, and slowed down her responses as she neared the middle, seeming to be unsure of the last callapo preferer (see Movie S3). This uncertainty could have been because she had made a grouping error or because she had not spatially separated the two groups. On the very next trial, AS correctly grouped the cards by preference and, strikingly, spatially separated the two groups into two distinct lines with a gap in between (Fig. 5, middle row, trial 4). Interestingly, this shape, “two adjacent lines,” did not appear on the first trial for any participant in the Preference condition nor on any trial in the other conditions. After the first trial, it was used by AS and two other participants, all of whom were in the Preference condition and had first grouped (or attempted to group) using a line, then on the subsequent trial added space between the groups, resulting in this “two adjacent lines” shape. Further, on this trial AS answered the first memory question partially correct, pointing to all the cards that had that preference (i.e., all the cards that preferred going to San Borja), and answered the second question fully correctly (i.e., the two males who preferred going to Yucumo). We interpret this trajectory of performance as suggesting that AS may have been discovering and fine-tuning this spatial strategy over the trials—even without feedback.

JM was approximately 50 years of age. His performance on the task was more difficult to interpret. On the first trial, he placed the cards in sequential order in a circular shape (Fig. 5, bottom row). He was the only participant in the experiment to create this shape, though others did in piloting. In the second trial, he seemed to start grouping cards by preference for the first four cards, but then it was unclear to the experimenter what he was doing after. After the last card was handed out, he moved the cards to be in more of a circle before the photo was taken (Movie S4). Before moving them, they had been in more of a line. He did not answer any of the memory questions accurately on the first two trials. On the third trial, he correctly grouped the cards by preference, still making a circular shape. Further, JM answered the memory questions partially correctly, accurately responding to the preference part of each question by pointing to the four cards on the left to answer “all the children that preferred canoe” and the four cards on the right to answer, “all the adults that preferred callapo.” This response pattern suggests that the grouping was intentional and useful to him. However, on the fourth trial, JM did not use this strategy again. Here, he began by placing cards sequentially in a rounded shape, but did not place the last three sequentially, and his question responses were not accurate. We interpret this trajectory of performance to suggest that JM may have also been discovering this spatial strategy while engaging with the task, and that, like many new strategies, it may not have been stable yet.

Note that participants’ initials have been changed to maintain their anonymity.

Order condition participants who reported no schooling. There were four participants in the Order condition who reported no schooling. These participants were not literate in either Tsimane’ or Spanish. NV was 53 years old, AC was around 65 (Movie S5), FC was approximately 30 (Movie S6), and CT was 29 (NV and CT asked to not be video recorded). The participants approached the task in different ways. NV and CT did not sequentially order their cards on the first trial, whereas AC and FC did (Fig. S5). NV started sequentially ordering on T2 and CT started on T3 (Fig. S5). All four created different shapes to represent sequential order and placed cards in different directions. For example, NV made two rows and the sequence snaked through the shape. AC made two columns, starting at the top for both columns on T1 and 2, and switching to the sequence snaking through the shape on T3 and 4. FC placed the cards left to

right and put the last one or two on a row above the first row. She put the cards face-down in T1 and 2, and then switched to placing them face up on T3. CT created lines and started by placing cards with a leftward bias on T1, switching to a more rightward bias on T2, and on T3 and 4 ordered the cards, placing them right to left. Note that participants' initials have been changed to maintain their anonymity.

χ^2 Test of Independence results showed same results as Fisher's Exact Test

As explained in the Data Analysis section, in many cases a χ^2 Test of Independence could not be used because at least one of the expected value cells was less than 5. Therefore, for consistency, Fisher's Exact Test was used throughout. However, in some cases, a χ^2 test could be used. Here, we report the results of the χ^2 tests to show that the results show the same pattern as Fisher's Exact Test.

- Significantly more children and adults sequentially organized their card layouts in the Order condition than in the Control condition (children: $\chi^2(1, N = 31) = 10.10$, $BC-p = .004$; adults: $\chi^2(1, N = 45) = 6.64$, $BC-p = .030$). Fisher's Exact Test results (see Table S4): children: $OR = 26.82$, 95% CI [2.76, 1403.11], $BC-p = .002$; adults: $OR = 6.54$, 95% CI [1.48, 35.80], $BC-p = .017$
- Significantly more adults sequentially organized their cards in the Order condition than in the Preference condition ($\chi^2(1, N = 32) = 6.06$, $BC-p = .041$). Fisher's Exact Test results: $OR = 8.22$, 95% CI [1.45, 60.27], $BC-p = .035$
- Adults in the Order condition created lines significantly more often than adults in the Control condition ($\chi^2(1, N = 46) = 12.50$, $BC-p = .001$), but there was no difference between Order and Preference conditions ($\chi^2(1, N = 32) = 2.01$, $BC-p = .469$). Fisher's Exact Test results: Order versus Control: $OR = 14.63$, 95% CI [2.79, 108.35], $BC-p < .001$; Order versus Preference: $OR = 3.51$, 95% CI [0.69, 20.43], $BC-p = .467$
- Children created lines equally often across all three conditions ($\chi^2(2, N = 46) = 1.37$, $p = .504$). Fisher's Exact Test results: $p = .599$

Supplemental Figures

Trial	Sequential order								Preferences	Preference groups
	1	2	3	4	5	6	7	8		
1									plantain or coconut	
2									fishing or harvesting rice	
3									canoe or "callapo"	
4									San Borja or Yucumo	

Fig. S1. Face stimuli for each trial. The “Sequential order” columns show the order the cards were handed out, which was the same across all three conditions and was the relevant to-be-remembered in the Order condition. The “Preferences” column lists the two possible preferences, and the “Preference groups” column shows how the cards were organized by preference. The preferences on each trial were from different categories. Trial 1: food preference, Trial 2: work activity preference, Trial 3: boat preference, Trial 4: destination preference (the two closest cities to the Tsimane’ communities). Face images were adapted from [Generated.Photos](https://www.generated.photos/).

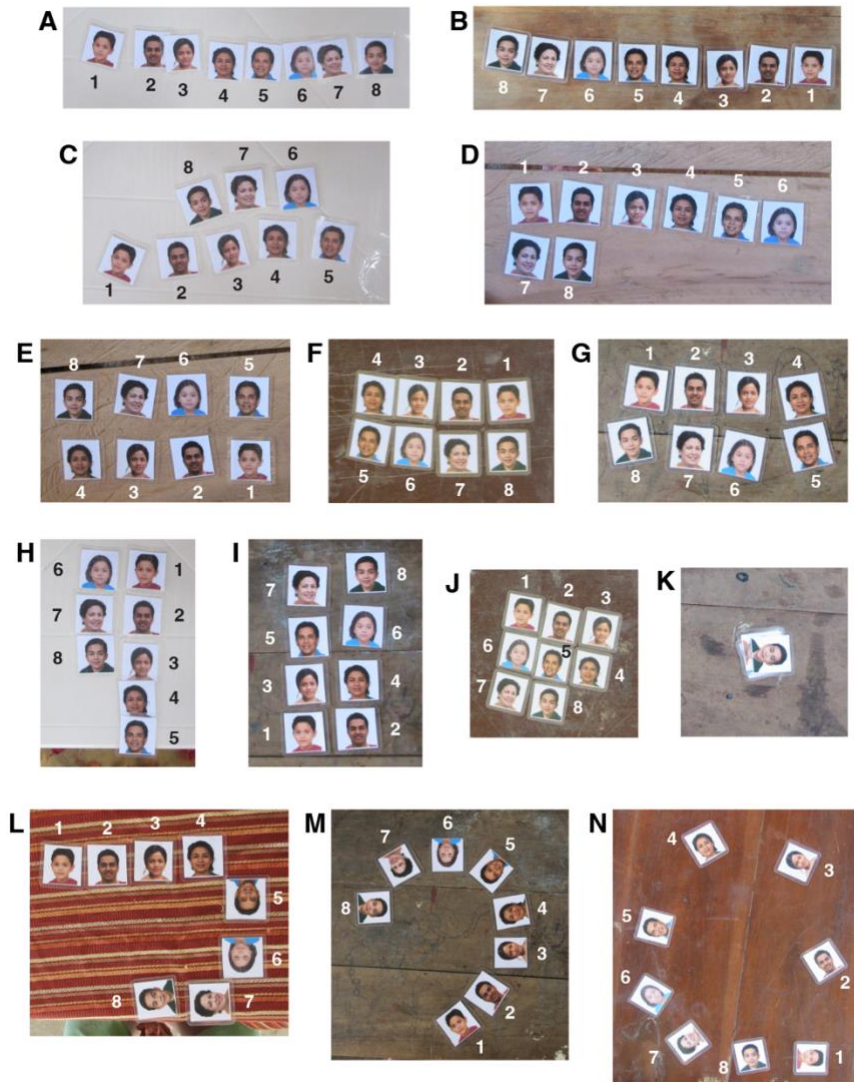


Fig. S2. Annotated images of sequentially ordered arrangements. Resulting card layouts from 14 different participants across conditions who sequentially ordered the cards on the first trial. The number annotations show the order that the cards were distributed. These examples reflect the variability in the shapes used to capture sequential order, as well as the variability in the directionality of card placement, even when the resulting shape created was the same (e.g., E-G). It is also worth noting that while four participants made sequentially ordered square grids (J) in the Control condition, this shape was never used to represent sequential order in the Order condition. In (K), the organization cannot be coded from the image, but was coded from the video. See Fig. S8 for original uncropped and unannotated images.

Order Condition									
Shape Name	line	line + extra		rectangle		"snake"	round (open)	1 pile	two clusters
Example Schematic									
Count of users on T1 (out of 24 who sequentially ordered)	14	3	1	2	1	1	1	1	0
Count of unique users across all trials	17	7	2	3	2	2	1	1	1

Preference Condition									
Shape Name	line	two clusters			rectangle		line + extra	round (closed)	unknown
Example Schematic									
Count of users on T1 (out of 11 who grouped)	4	3	1	0	0	1	0	1	1
Count of unique users across all trials	5	5	3	2 [‡]	1	5	2	1	1

Control Condition											
Shape Name	line	unknown / "random"		square	rectangle		line + extra	"snake"	1 pile	3 piles	two clusters
Example Schematic											
Count of users on T1 (out of 50)	15	5	4	8	7	1	4	1	2	2	1

Fig. S3. Distribution of shapes used across conditions. For Order and Preference conditions, the shapes shown are those created when organizing the cards to represent the relevant information (i.e., sequential for Order and grouped for Preference). "Count of users on T1" is the number of participants who created a given shape on T1 to represent the condition-relevant information. "Count of unique users across all trials" is the total number of participants who used a given shape on any trial to represent the condition-relevant information. Note that participants could have used one shape on one trial, and a different shape on a subsequent trial, so this row does not add up to the total number of participants. The Control condition shows the total number of participants who created each given shape on the first trial since there was no specified information to represent.

‡ There was a third participant who created this shape. She used a line to group on T2, and then on the next two trials she created this shape but made a grouping error both times. However, because the organization was coded as "attempted group" and this table only counts trials in which the organization was coded as grouped, this instance is not counted in the table.

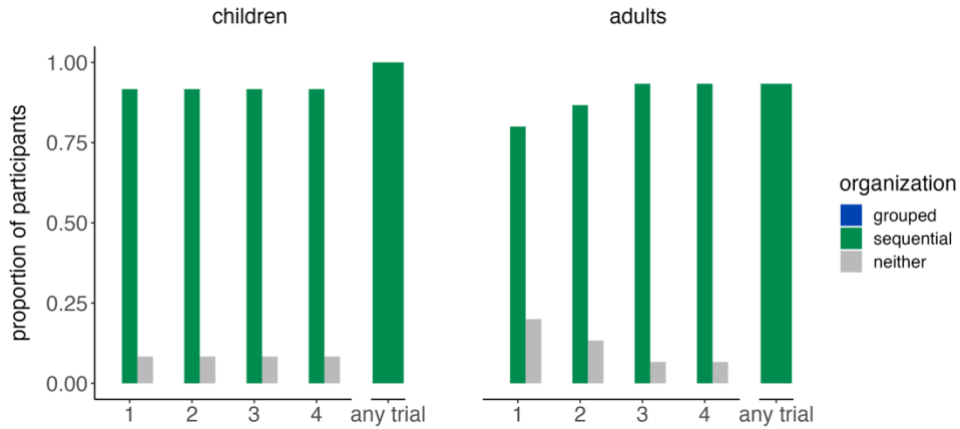


Fig. S4. Spatial organization across all trials in Order condition. Proportion of participants on each trial of the Order condition who used each organization: grouped, sequentially ordered, or neither. The last bar represents the proportion of participants who correctly sequentially ordered on at least one of the four trials.



Fig. S5. Annotated trial images of the four Order condition participants who reported no schooling. The number annotations show the order that the cards were distributed. Note that for participant NV, the experimenter made an error ordering the cards before the trial, swapping 3 and 7 in T1-3, so the images for T1-3 are annotated with the numbers indicating the order that the cards were handed to this particular participant. See Fig. S9 for original uncropped and unannotated images.

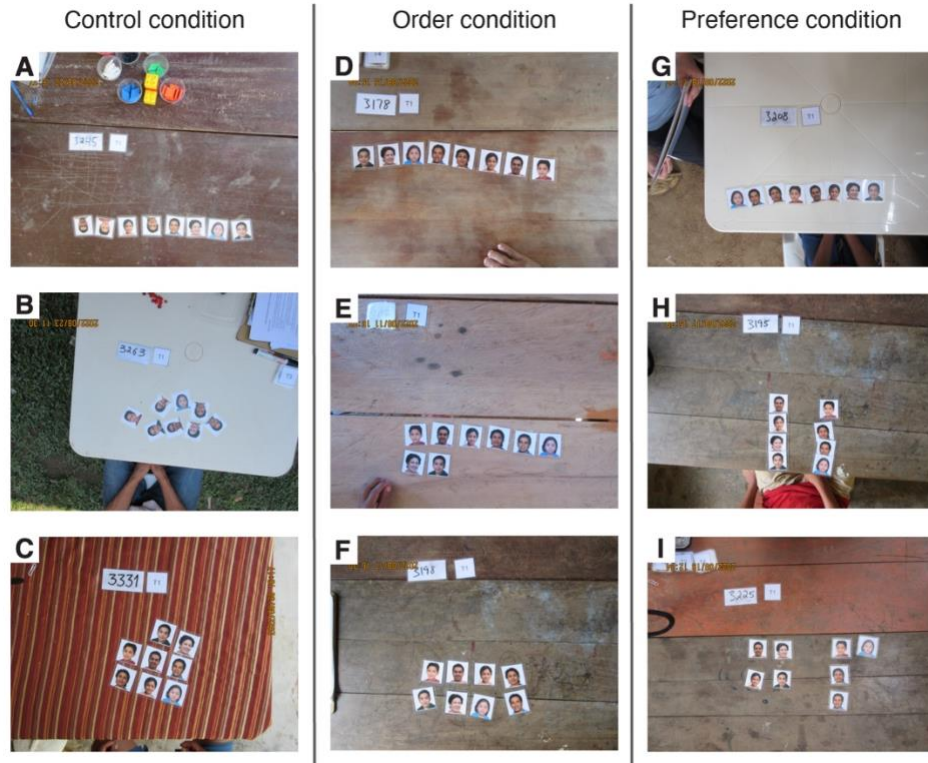


Fig. S6. Original images from Fig. 2, not cropped or annotated.

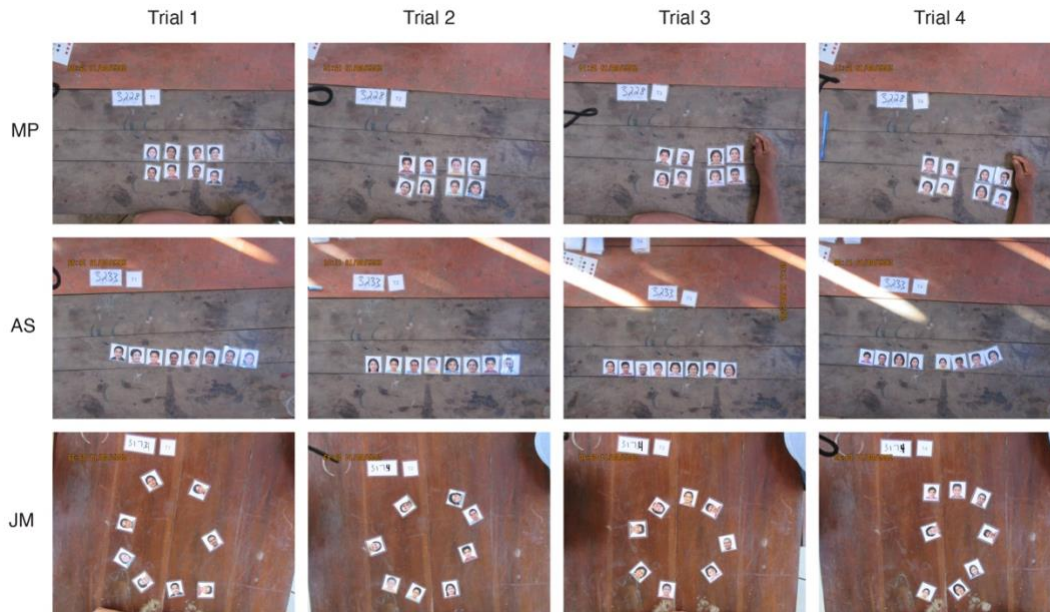


Fig. S7. Original images from Fig. 5, not cropped or annotated.

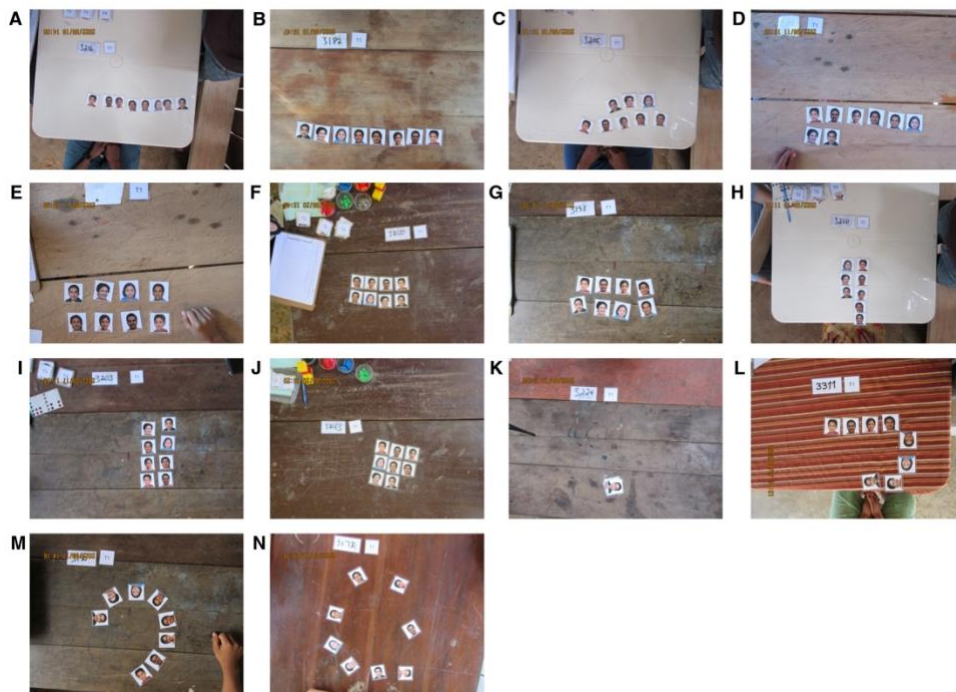


Fig. S8. Original images from Fig. Fig. S2, not cropped or annotated.

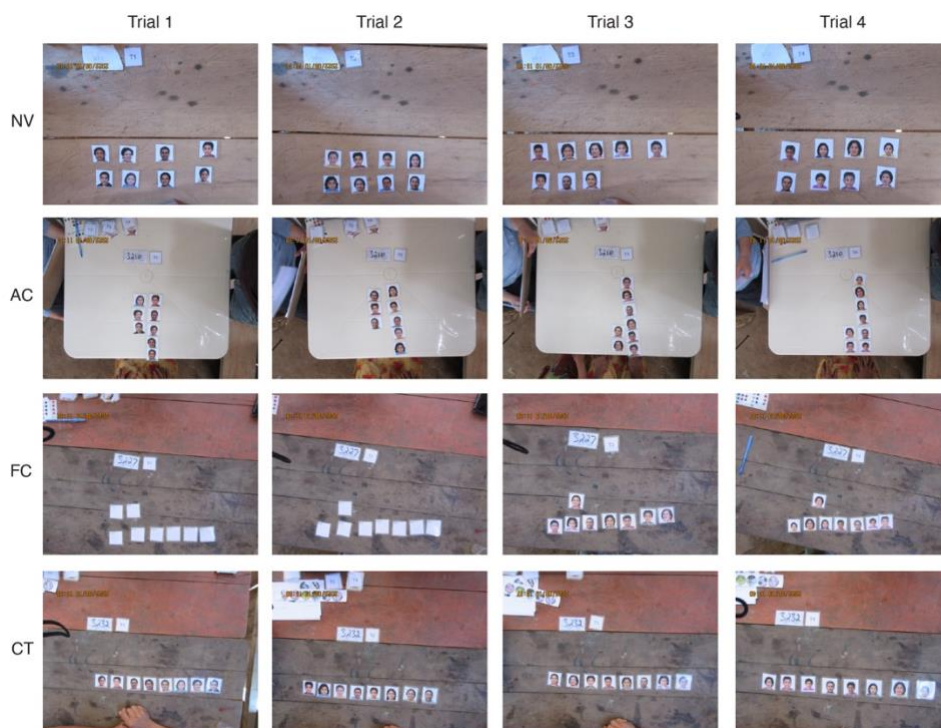


Fig. S9. Original images from Fig. Fig. S5, not cropped or annotated.

Supplemental Tables

Table S1. Sample size and demographic information of participants included in analysis

Variable	Children (n = 46)	Adults (n = 61)
Age	11.39 [7.50, 14.80], SD = 1.84	36.19 [16.10, 90.00], SD = 17.49
Schooling	4.54 [1, 9], SD = 1.96	4.36 [0, 13], SD = 3.58
Literacy	1.87, SD = 2.74	5.05, SD = 3.70

Note: For age and schooling: Mean [Range], SD in years; Literacy: Mean, SD, range is 0 to 8 by definition.

Table S2. Number of participants in each condition by age group

Age Group	Control	Order	Preference	n
Children	21	12	13	46
Adults	29	17	15	61
Total	50	29	28	107

Table S3. Memory test questions for each experimental condition and trial. The numbers in the “Answer” columns refer to the position of the card in the sequence for that trial. See Fig. S1 for the card position numbers and preference groups.

Trial	Conjunctive dimension	Order		Preference	
		Question	Answer	Question	Answer
1	Age group	Who was the first child to arrive?	1	Who are all the adults who prefer plantain?	4 and 5
		Who was the last adult to arrive?	7	Who are all the children who prefer coconut?	3 and 8
2	Gender	Who was the first male to arrive?	3	Who are all the males who prefer fishing?	4 and 8
		Who was the last female to arrive?	6	Who are all the females who prefer harvesting rice?	1 and 6
3	Age group	Who was the first adult to arrive?	3	Who are all the children who prefer going in canoe?	1 and 6
		Who was the last child to arrive?	8	Who are all the adults who prefer going in “callapo”?	3 and 5
4	Gender	Who was the first female to arrive?	1	Who are all the females who prefer going to San Borja?	2 and 3
		Who was the last male to arrive?	7	Who are all the males who prefer going to Yucumo?	4 and 7

Table S4. Results of four Fisher’s Exact Tests, comparing sequential ordering in the Order condition to (1) sequential ordering in the Control and (2) sequential ordering in the Preference condition. Tests were run separately for children and adults. In the Order condition, 11/12 children and 13/17 adults sequentially ordered.

Condition	Age Group	Proportion Sequential	Odds Ratio	95% CI	Bonferroni-Corrected p-value
Control	Children	5/19	26.82	[2.76, 1,403.11]	.002
	Adults	9/28	6.54	[1.48, 35.80]	.017
Preference	Children	4/13	21.07	[1.95, 1,169.85]	.011
	Adults	4/15	8.22	[1.45, 60.27]	.035

Table S5. Beta coefficients for a logistic mixed effects model predicting memory question accuracy in the Preference condition from question number, trial number, and organization (grouped or not grouped).

Term	<i>b</i>	95% CI	<i>z</i>	<i>p</i>
Intercept	-4.05	[-5.65, -2.44]	-4.94	< .001
Question	0.15	[-0.61, 0.90]	0.38	.701
Trial	0.52	[0.15, 0.89]	2.77	.006
Grouped	3.45	[2.21, 4.69]	5.47	< .001

Table S6. Beta coefficients for a logistic mixed effects model predicting memory question accuracy in the Order condition from question number, trial number, and organization (sequential or not sequential).

Term	<i>b</i>	95% CI	<i>z</i>	<i>p</i>
Intercept	1.00	[-0.17, 2.18]	1.67	.095
Question	-1.50	[-2.12, -0.88]	-4.74	< .001
Trial	0.15	[-0.12, 0.42]	1.09	.274
Sequential	-0.13	[-1.18, 0.92]	-0.24	.810

Table S7. Number of children and adults that used each organization on the first trial by condition. These counts are visualized as within-condition proportions in Fig. 3A.

Age group	Condition	Organization			n
		Grouped	Sequential	Neither	
Children	Control	0	5	14	19
	Order	0	11	1	12
	Preference	2	4	7	13
Adults	Control	0	9	19	28
	Order	0	13	4	17
	Preference	9	4	2	15

Note: In the Control condition the layouts from 2 children and 1 adult could not be coded for organization from the picture due to the shape (see sections on coding organization and data cleaning), resulting in the *n* in this table for the Control condition as opposed to the full sample reported in Table S2.

Table S8. Number of children and adults that used lines versus all other shapes on the first trial by condition. The proportion that used lines is the height of the black bars in Fig. 3B.

Age Group	Condition	Shape		n
		Line	Other Shape	
Children	Control	12	9	21
	Order	5	7	12
	Preference	5	8	13
Adults	Control	3	26	29
	Order	11	6	17
	Preference	5	10	15

Table S9. Organization of the lines that children and adults created on the first trial by condition. The heights of the colored stacked bars in Fig. 3B come from these counts divided by the total number of participants in the condition (n).

Age Group	Condition	Organization			Total Lines	n
		Grouped	Sequential	Neither		
Children	Control	0	4	8	12	21
	Order	0	5	0	5	12
	Preference	1	2	2	5	13
Adults	Control	0	2	1	3	29
	Order	0	9	2	11	17
	Preference	3	1	1	5	15

Table S10. Organizations created by children and adults in the Preference condition over the four trials. The row “any trial” shows the number of participants who grouped correctly on at least one trial. The counts in this table are visualized as proportions out of the total number of participants in this condition (13 children and 15 adults) in Fig. 4.

Age Group	Trial	Organization			
		Grouped	Attempted	Sequential	Neither
Children	1	2	0	4	7
	2	7	0	4	2
	3	5	1	5	2
	4	7	1	4	1
	any trial	8	---	---	---
Adults	1	9	0	4	2
	2	5	3	3	4
	3	7	4	2	2
	4	8	4	2	1
	any trial	13	---	---	---

Table S11. Organizations created by children and adults in the Order condition over the four trials. The row “any trial” shows the number of participants who sequentially ordered on at least one trial. The counts in this table are visualized as proportions out of the total number of participants in this condition (12 children and 17 adults) in Fig. S4.

Age Group	Trial	Organization	
		Sequential	Neither
Children	1	11	1
	2	11	1
	3	11	1
	4	11	1
	any trial	12	---
Adults	1	12	3
	2	13	2
	3	14	1
	4	14	1
	any trial	14	---

Note: Because no participants grouped in this condition, this column was not included in the table.

Supplementary Movies

Movie S1. All images from the first trial, organized by condition. Hosted on OSF: <https://osf.io/rgje7>

Movie S2. Video of Preference condition participant MP (age ~40) completing the four trials of the task. She reported no schooling and was not literate. MP grouped the cards by preference on the first trial. The video has English subtitles, and the trials are annotated to show which cards belong to each preference. Hosted on OSF: <https://osf.io/8hzuv>

Movie S3. Video of Preference condition participant AS (age 58) completing the four trials of the task. She reported no schooling and was not literate. AS did not group by preference on the first trial and answered both memory questions incorrectly. She changes her strategy and on T4 she correctly grouped with space separating the groups. The video has English subtitles, and the trials are annotated to show which cards belong to each preference. Hosted on OSF: <https://osf.io/fx52b>

Movie S4. Video of Preference condition participant JM (age ~50) completing the four trials of the task. He reported no schooling and was not literate. JM did not group by preference on the first trial but does so on T3. The video has English subtitles, and the trials are annotated to show which cards belong to each preference. Hosted on OSF: <https://osf.io/4ugc6>

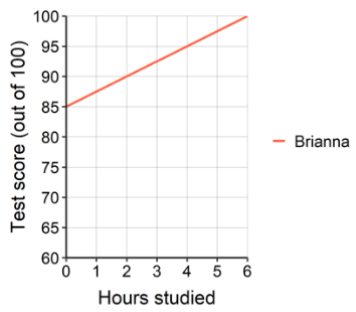
Movie S5. Video of a participant AC (age ~65), who was in the Order condition, completing the four trials of the task. She reported no schooling and was not literate. AC sequentially ordered the cards on the first trial. The video has English subtitles. Hosted on OSF: <https://osf.io/fxw84>

Movie S6. Video of a participant FC (age ~30), who was in the Order condition, completing the four trials of the task. She reported no schooling and was not literate. FC sequentially ordered the cards on the first trial. The video has English subtitles. Hosted on OSF: <https://osf.io/y6tp7>

Appendix C: Supplemental Materials for Chapter 4

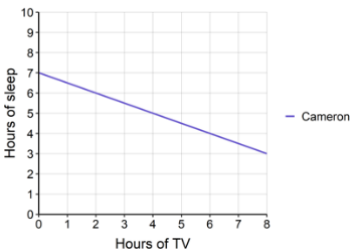
The following sections show the graphs and questions used for each task. The question type, either multiple choice (MC) or open-ended (OE), is also included in parentheses next to the question number.

Basic graph knowledge assessment questions

Graph	Question (Type)	Question
	Q1 (MC)	What variable is on the x-axis?
	Q2 (MC)	What variable is on the y-axis?
	Q3 (MC)	If Brianna studied for 2 hours, what would her score on the test be?
	Q4 (MC)	What is the value of y when $x = 4$?
	Q5 (MC)	If Brianna wants to score a 100 on the test, how many hours does she need to study?
	Q6 (MC)	What is the value of x when $y = 85$?

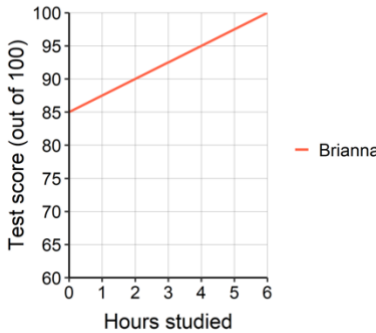
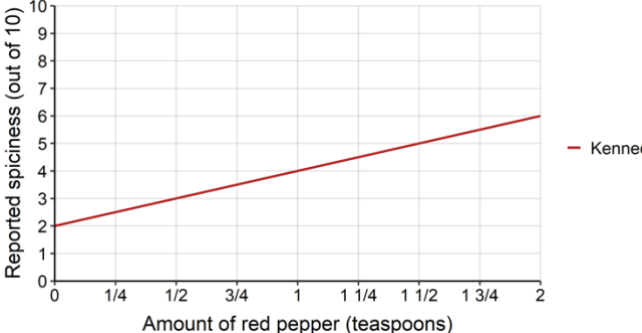
Extra graph practice

For students who got less than 5 of the questions correct on the basic graph knowledge assessment.

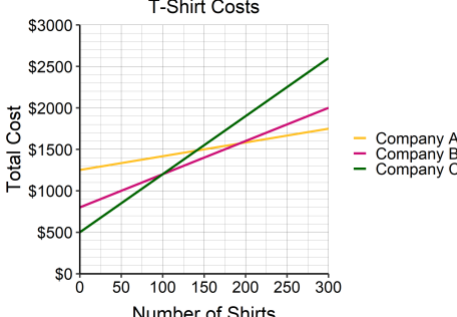
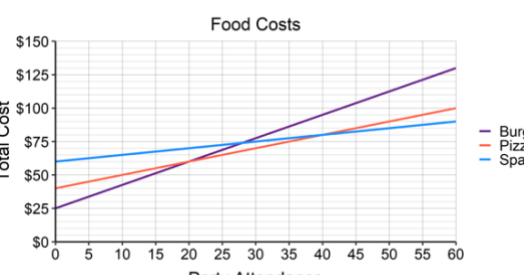
Graph	Question (Type)	Question
	Q1 (MC)	If Cameron watched 8 hours of TV, how many hours will he sleep?
	Q2 (MC)	When $x = 0$, what is the value of y ?
	Q3 (MC)	When Cameron sleeps for 5 hours, how many hours of TV did he watch according to this graph?
	Q4 (MC)	What is the value of x when $y = 6$?

Pre- and post-test

Part 1: y-intercept and slope

Question (Type)	Pre-test	Post-test
Graph	 <p>A line graph with 'Hours studied' on the x-axis (0 to 6) and 'Test score (out of 100)' on the y-axis (60 to 100). A red line labeled 'Brianna' starts at (0, 85) and passes through (6, 100).</p>	 <p>A line graph with 'Amount of red pepper (teaspoons)' on the x-axis (0 to 2) and 'Reported spiciness (out of 10)' on the y-axis (0 to 10). A red line labeled 'Kennedy' starts at (0, 2) and passes through (2, 6).</p>
Cover Story	NA	<p>Kennedy loves to cook chili. Each time she makes it, she adds different amounts of red pepper powder and records how spicy it tastes on a scale of 0 to 10, where 10 is extremely spicy.</p> <p>Here's a graph showing the amount of red pepper added on the x-axis and the reported spiciness on the y-axis.</p>
Q1 (click)	Click on the y-intercept using your mouse.	Click on the y-intercept using your mouse.
Q2 (OE)	What does the y-intercept on this graph tell you?	What does the y-intercept on this graph tell you?
Q3 (OE)	Describe the slope on this graph. What does it tell you?	Describe the slope on this graph. What does it tell you?

Part 2: problem solving transfer task

Question (Type)	Pre-test	Post-test
Graph	 <p>The graph shows three linear functions representing the total cost of t-shirts from three different companies. The x-axis is 'Number of Shirts' (0 to 300) and the y-axis is 'Total Cost' (\$0 to \$3000). Company A (yellow) starts at approximately \$1250 for 0 shirts and increases to about \$1750 for 300 shirts. Company B (pink) starts at approximately \$800 for 0 shirts and increases to about \$2000 for 300 shirts. Company C (green) starts at \$500 for 0 shirts and increases to about \$2600 for 300 shirts.</p>	 <p>The graph shows three linear functions representing the total cost of food for a party based on the number of attendees. The x-axis is 'Party Attendance' (0 to 60) and the y-axis is 'Total Cost' (\$0 to \$150). Burgers (purple) starts at approximately \$25 for 0 attendees and increases to about \$130 for 60 attendees. Pizza (orange) starts at approximately \$40 for 0 attendees and increases to about \$100 for 60 attendees. Spaghetti (blue) starts at approximately \$60 for 0 attendees and increases to about \$90 for 60 attendees.</p>
Cover Story	<p>The student council at Sunnyside Middle school is planning to sell school t-shirts to the 8th graders.</p> <p>They are deciding between three different companies, and they need to figure out which is the cheapest option.</p> <p>Your mission is to help the student council decide which t-shirt company is the cheapest option.</p> <p>Here's a graph showing relationship between the number of t-shirts purchased and the total cost in dollars for the three different t-shirt companies.</p>	<p>You are hosting a birthday party for your best friend! You're trying to decide what food to serve. You have three ideas in mind for what you could serve: burgers, pizza, or spaghetti.</p> <p>You want to pick the food that will cost you the least amount of money.</p> <p>In order to make your decision, you look at a graph that shows you the relationship between the number of friends coming to the party and the total cost for the three different food options.</p>
Describe prompt* (OE)	Describe to student council what is going on in the graph.	Describe to your friend what is going on in the graph.
Q1 (OE)	There are 300 students in the 8 th grade class, but the student council does not know yet how many students will buy a t-shirt.	<p>You invited 60 friends to the party, but you do not know yet how many will actually end up coming.</p> <p>Make a recommendation about which food option you should use for the party.</p>

	<p>Make a recommendation to student council about which company they should use to make the t-shirts.</p> <p>In your response, explain how you made your choice.</p>	<p>In your response, explain how you made your choice.</p>
Q2 (OE)	<p>Under which circumstances, if ever, would you pick each of the three companies?</p> <p>Explain your reasoning, and be sure to include what about the graph made you give this answer.</p>	<p>Under which circumstances, if ever, would you pick each of the three food options?</p> <p>Explain your reasoning, and be sure to include what about the graph made you give this answer.</p>
Q3 (OE)	<p>When, if ever, would buying from Company A be the cheapest option?</p>	<p>When, if ever, would ordering burgers be the cheapest option?</p>
Q4 (OE)	<p>When, if ever, would buying from Company B be the cheapest option?</p>	<p>When, if ever, would ordering pizza be the cheapest option?</p>
Q5 (MC)	<p>Which company has the lowest starting cost?</p> <p>Options: A, B, C, You can't tell from this graph, I don't know</p>	<p>Which food option has the lowest starting cost?</p> <p>Options: Burgers, Pizza, Spaghetti, You can't tell from this graph, I don't know</p>
Q6 (MC)	<p>Which company has the lowest cost for each additional t-shirt purchased?</p> <p>Options: A, B, C, You can't tell from this graph, I don't know</p>	<p>Which food option has the lowest cost for each additional friend who comes to the party?</p> <p>Options: Burgers, Pizza, Spaghetti, You can't tell from this graph, I don't know</p>

* This question was exploratory, as described in the preregistration. It was not scored for this study or used in the analysis.

These questions are organized from most broad to most specific with the goal of being able to observe what kinds of answers students give spontaneously (e.g., what kind of reasoning they engage in), before seeing how students answer when prompted for more specific information and for a more specific style of reasoning. The goal was that this style of scaffolded questions would

make a more sensitive measure, with the highest scores for those who generate complex reasoning responses spontaneously from the beginning, with the next highest scores for those who reason more complexly when prompted, and finally those who have not reasoned complexly. See the scoring guide for additional details.

Interrater reliability for open-ended questions

Interrater reliability was calculated using Pearson’s correlation coefficient (r) since there were two raters and the scores were continuous. Overall agreement across all open-ended questions was high ($r = .94$).

Table S1. Interrater reliability for the open-ended questions in the first part of the pre-test measured by Pearson’s r .

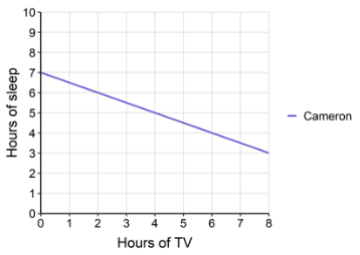
Question	Timepoint	Pearson’s r
Part 1: Q2	Pre	.92
	Post	.90
	Overall	.92
Part 1: Q3	Pre	.95
	Post	.90
	Overall	.93
Part 2: Q1	Pre	.84
	Post	.83
	Overall	.84
Part 2: Q2	Pre	.95
	Post	.99
	Overall	.97
Part 2: Q3	Pre	.95
	Post	.97
	Overall	.96
Part 2: Q4	Pre	.96
	Post	.96
	Overall	.96
All responses		.94

Lesson

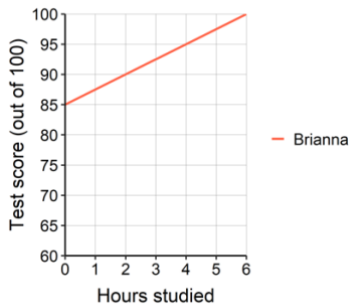
Instructional Block

y-intercept

Question (Type)	Graph	Visual Features Lesson	Relational Reasoning Lesson
instructional video	<p>Test score (out of 100)</p> <p>Hours studied</p> <p>— Brianna</p>	<p>The y-intercept is the value of y when $x = 0$.</p> <p>In other words, it is where the line intersects with the y-axis.</p>	<p>In other words, this value can be thought of as a “baseline” or starting point.</p>
practice Q1 (MC)	<p>Test score (out of 100)</p> <p>Hours studied</p> <p>— Brianna</p>	<p>What is the y-intercept of Brianna’s line? In other words, when $x = 0$?</p>	<p>What would Brianna score on the test if she decided not to study at all? In other words, when hours studied = 0.</p>
practice Q2 (MC)	<p>Test score (out of 100)</p> <p>Hours studied</p> <p>— Brianna — James</p>	<p>What is James’s y-intercept?</p>	<p>What would James score on the test if he decided not to study at all?</p>
practice Q3 (MC)	<p>Test score (out of 100)</p> <p>Hours studied</p> <p>— Brianna — James</p>	<p>Which student has the higher y-intercept?</p>	<p>Which student would score higher on the test without studying?</p>

practice Q4 (MC)		What is the y-intercept of Cameron's line?	How many hours of sleep would Cameron get if he didn't watch any TV in a day?
------------------	---	--	---

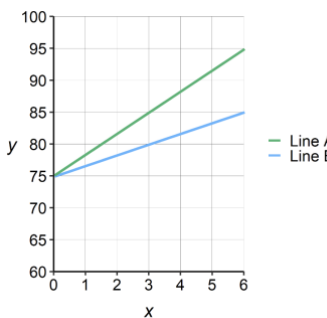
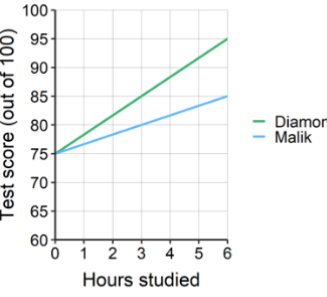
Slope direction

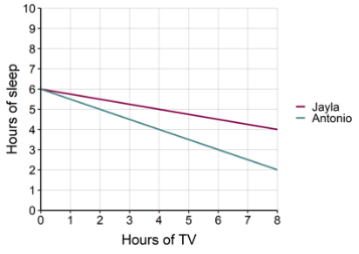
Question (Type)	Graph	Visual Features Lesson	Relational Reasoning Lesson
instructional video		<p>The slope describes both the direction and the steepness of the line.</p> <p>The direction of the line can be positive, negative, or 0.</p>	<p>The slope of a line represents the nature of the relationship between x and y.</p> <p>This relationship can be positive, negative, or 0.</p>
	A positive slope means that...		
	<p>... the line is going <i>up</i> from left to right.</p> <p>For example, the slope of Brianna's line is positive since it is going up from left to right.</p>	<p>... as x increases, the value of y also increases.</p> <p>For example, the slope of Brianna's line is positive since as she studies more for the test, her test score also increases.</p>	
A negative slope means that...			

		<p>... the line is going <i>down</i> from left to right.</p> <p>For example, the slope of Cameron's line is negative since it is going down from left to right.</p>	<p>... as x increases, the value of y decreases.</p> <p>For example, the slope of Cameron's line is negative since when he watches <i>more</i> hours of TV in a day, he sleeps <i>fewer</i> hours at night.</p>
		<p>Here, the slope of Amari's line is 0 since...</p>	
		<p>... the line is flat as it goes from left to right.</p>	<p>... when she watches <i>more</i> hours of TV in a day, her sleep at night doesn't change, it stays the same.</p>
<p>practice Q1 (MC)</p>		<p>The line is going down from left to right. The slope of this line is _____.</p> <p>Options: positive, negative, 0</p>	<p>As x increases, y decreases. The slope of this line is _____.</p> <p>Options: positive, negative, 0</p>
<p>practice Q2 (MC)</p>	<p>[No graph presented for these questions]</p>	<p>The line is flat from left to right. The slope of this line is _____.</p> <p>Options: positive, negative, 0</p>	<p>As x increases, y stays the same. The slope of this line is _____.</p> <p>Options: positive, negative, 0</p>
<p>practice Q3 (MC)</p>		<p>The line is going up from left to right. The slope of this line is _____.</p> <p>Options: positive, negative, 0</p>	<p>As x increases, y increases. The slope of this line is _____.</p> <p>Options: positive, negative, 0</p>

Note that the way that graphs were animated in the instructional videos also differed by condition. See the videos posted on OSF to view the animations in action.

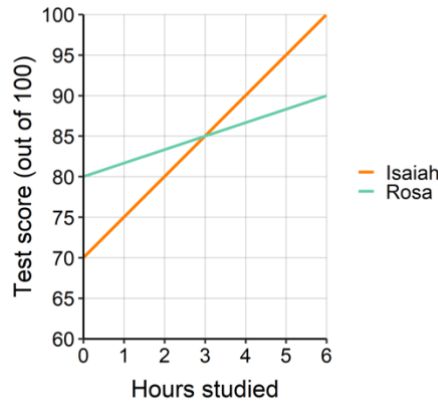
Slope steepness

Question (Type)	Graph	Visual Features Lesson	Relational Reasoning Lesson
		<p>In addition to the slope being positive, negative, or 0, the slope can be steeper or shallower depending on...</p>	
	<p>... the angle of the line.</p> <p>Steeper lines go up or down <i>faster</i> than shallower lines.</p>	<p>... how much y changes for each increase in x.</p> <p>As x increases, y changes (either increases or decreases) <i>more</i> for a steeper line than for a shallower line.</p>	
instructional video		<p>Let's use the graph of hours studied and test score to take a closer look. This time, we are going to be look at the data for two students, Diamond and Malik.</p> <p>In this graph, Diamond and Malik have the same y-intercept, which means that...</p>	
		<p>... when x is equal to 0, y = 75 for both students.</p>	<p>... if they both study 0 hours for a test, they would get the same score on the test, a 75.</p>
		<p>Even though they both have positive slopes, Diamond's like has a <i>steeper</i> positive slope.</p>	
		<p>This means that as you go from left to right, Diamond's line goes up faster than Malik's line.</p>	<p>This means that for each hour the students study, Diamond's test score improves <i>more</i> from that hour of</p>

			studying that Malik's score.
practice Q1 (MC)	 <p>The graph shows two downward-sloping lines. Jayla's line starts at 6 hours of sleep for 0 hours of TV and decreases to 4 hours of sleep for 8 hours of TV. Antonio's line starts at 6 hours of sleep for 0 hours of TV and decreases to 2 hours of sleep for 8 hours of TV. Jayla's line has a shallower slope.</p>	Which person's line has the shallower slope?	Which person's sleep is <i>less</i> affected by each additional hour of TV watched?

Practice Block

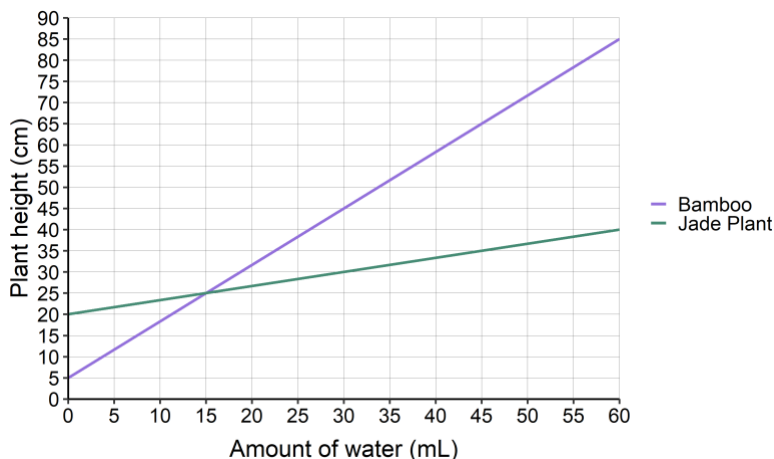
Graph 1 (Day 1)



Question (type)	Visual Features Lesson	Relational Reasoning Lesson
Cover story	Here's a graph showing hours studied on the x-axis and test score on the y-axis with lines for two new students, Isaiah and Rosa.	Here's a graph showing the relationship between hours studied and test score for two new students, Isaiah and Rosa.
Q1 (OE)	How would you explain to a friend what you see in this graph? <i>Feel free to use any terms you've learned.</i>	A) What are some similarities, if any, in the relationship between hours studied and test score for Isaiah and Rosa? B) What are some differences, if any, in the relationship between hours and test score for Isaiah and Rosa? <i>Feel free to use any terms you've learned.</i>
Q2 (MC)	Which student has the greater y-intercept?	If Isaiah and Rosa decided not to study for the test at all, which student would score higher on the test?
Q3 (MC)	Which student's line has a steeper slope?	Which student's test score would benefit more from studying for more hours?
Q4 (MC)	Which student has the greater y-value when $x = 2$?	Which student would score higher on the test if they both had less than 3 hours to study?

Q5 (VF: MC; RR: OE)	Which student has the greater y-value when $x = 6$?	For what amounts of study time would Isaiah score higher on the test than Rosa?
---------------------------	--	---

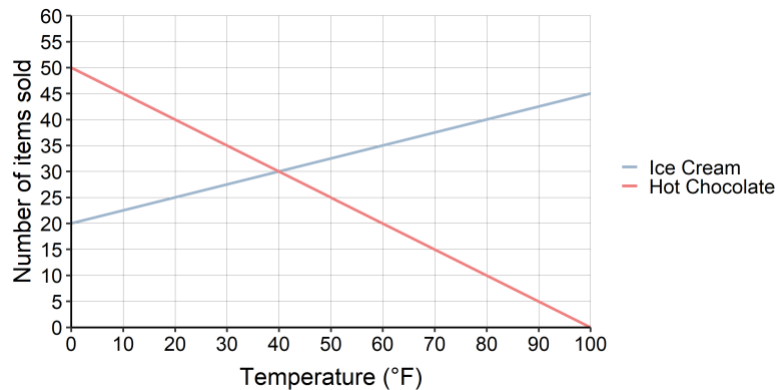
Graph 2 (Day 1)



Question (type)	Visual Features Lesson	Relational Reasoning Lesson
Cover story	Your friend is growing bamboo and jade plants. Here's a graph showing the amount of water given to the plants every day on the x-axis and how tall the plant will be after 2 months on the y-axis, with lines for the two plants.	Your friend is growing bamboo and jade plants, and wants to understand the relationship between the amount of water given to the plants every day and how tall they will be after 2 months.
Q1 (OE)	How would you explain to a friend what you see in this graph? <i>Feel free to use any terms you've learned.</i>	A) What are some similarities, if any, in the relationship between amount of water and plant height for bamboo and jade? B) What are some differences, if any, in the relationship between amount of water and plant height for bamboo and jade? <i>Feel free to use any terms you've learned.</i>
Q2 (MC)	Which plant has the lower y-intercept?	If no water is given to either plant, which plant would be shorter?

Q3 (MC)	Which plant's line has the steeper slope?	Which plant would grow more from being watered more?
Q4 (MC)	Which plant has the higher y-value when $x = 10$?	Which plant would be taller if they were both watered less than 15mL a day?
Q5 (VF: MC; RR: OE)	Which plant has the greater y-value when $x = 45$?	For what amounts of water would bamboo be taller than jade?

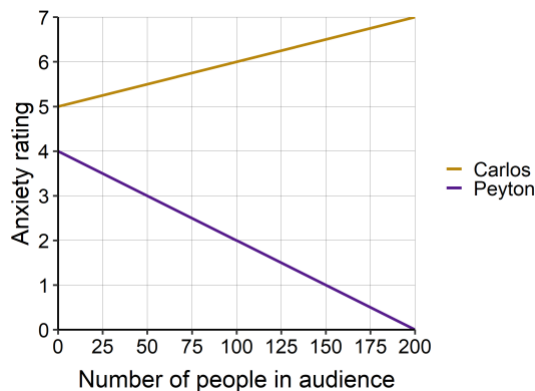
Graph 3 (Day 2)



Question (type)	Visual Features Lesson	Relational Reasoning Lesson
Cover story	Here's a graph showing temperature on the x-axis and number of items sold on the y-axis, with lines for two kinds of snack bar food.	Here's a graph showing the relationship between the temperature outside and the number of items sold at a snack bar for two kinds of food.
Q1 (OE)	How would you explain to a friend what you see in this graph? <i>Feel free to use any terms you've learned.</i>	A) What are some similarities, if any, in the relationship between temperature and number of items sold for ice cream and hot chocolate? B) What are some differences, if any, in the relationship between temperature and number of items sold for ice cream and hot chocolate? <i>Feel free to use any terms you've learned.</i>
Q2 (MC)	Which food has the greater y-intercept?	When it is 0 degrees outside, which item would sell more?

Q3 (A: MC, B: OE)	A) Which food's line has a steeper slope? B) Explain how you know.	A) Which food shows the bigger change in sales between 0F and 100F? B) Explain how you know.
Q4 (VF: MC; RR: OE)	Which food has the greater y-value when $x = 80$?	When would ice cream sales be more than hot chocolate sales?
Q5 (OE)	What would x be when $y = 30$ for both foods?	Would ice cream sales and hot chocolate sales ever be equal? If so, when?
Q6 (OE)	What would x be when $y = 25$ for both foods?	Imagine you are the store owner and you want to sell <i>at least</i> 25 ice creams and 25 hot chocolates in a day. For what temperatures would you want to open your store to accomplish this goal?

Graph 4 (Day 2)



Question (type)	Visual Features Lesson	Relational Reasoning Lesson
Cover story	Here's a graph showing the number of people in an audience on the x-axis and anxiety rating on the y-axis, with lines for two people who are in a musical.	Here's a graph showing the relationship between the number of people in an audience and the anxiety rating for two people who are in a musical.
Q1 (OE)	How would you explain to a friend what you see in this graph? <i>Feel free to use any terms you've learned.</i>	A) What are some similarities, if any, in the relationship between number of people in the audience and anxiety rating for Carlos and Payton?

		<p>B) What are some differences, if any, in the relationship between number of people in the audience and anxiety rating for Carlos and Payton?</p> <p><i>Feel free to use any terms you've learned.</i></p>
Q2 (MC)	Which person has the <i>lower</i> y-intercept?	If there were no people in the audience, which person would be <i>less</i> anxious?
Q3 (MC)	Which person's line has a positive slope?	Which person would get more anxious having more people in the audience?
Q4 (A: MC, B: OE)	<p>A) Which person's line has the steeper slope?</p> <p>B) Explain how you know.</p>	<p>A) Which person's anxiety rating changes the most as the number of people in the audience increases?</p> <p>B) Explain how you know.</p>
Q5 (VF: MC; RR: OE)	When $x = 100$, which person's y-value is greater?	For what numbers of people in the audience would Carlos be more anxious than Peyton?