7 Causal Reasoning and the Development of Metacognitive Thinking: Cross-Sectional and Longitudinal Investigations

Here in part II, we have charted the development of reasoning capacities related to scientific thinking, using a new diagnostic reasoning measure to begin to bridge the gap between children's early success at engaging in causal reasoning and their later difficulties with scientific thinking. We have already seen that causal complexity is a factor in this developmental trajectory. In the studies we have presented, children begin to make inferences about our reasoning task in a similar manner to adults at ages 7 to 8 years, but not earlier (chapter 5). Further, we saw that making this reasoning problem more contextualized by including more real-world science content makes it more difficult for children to solve (chapter 6). Although more work needs to be done on both of these topics to really settle how different types of causal structures and different aspects of a particular contextualization interact with children's scientific thinking abilities, here we turn to an exploration of a different variable that could affect this process: children's ability to understand and resolve disagreements.

This ability is part of children's developing metacognitive capacities, a suite of abilities that involve explicit reflection on the nature of beliefs and how they function. In general, children by about the age of 4 can understand that a particular belief can be false and that a single agent can hold a belief that does not match the truth of current reality (e.g., Flavell et al., 1992; Gopnik & Astington, 1988; Perner et al., 1987; Wellman & Liu, 2004). But there is much more to understanding beliefs than understanding that others' beliefs can be false.

A good example of the continued trajectory of children's understanding of beliefs can be seen in their performance on a battery of tasks aimed at probing

their *advanced theory of mind* abilities. Advanced theory of mind includes abilities such as recognizing the complexity of others' mental states (e.g., understanding the recursive nature of belief and its relation to other mental states; e.g., Eisbach, 2004; Lagattuta & Wellman, 2001; Perner & Wimmer, 1985), broad perspective-taking capacities (e.g., Carpendale & Chandler, 1996), and understanding and recognizing others' emotions (e.g., Baron-Cohen et al., 2001). This kind of mental-state knowledge continues to develop during the elementary school years (e.g., Osterhaus et al., 2016). Further, such understanding is composed of multiple facets, meaning that there is not a single developmental trajectory for advanced theory of mind.

One aspect of children's metacognitive understanding that we believe is critical for scientific reasoning is the ability to negotiate situations in which agents hold different beliefs, or situations in which agents' own beliefs need to change in light of new evidence—that is, cases of disagreement or belief conflict (Beck, Robinson et al., 2011; Heiphetz, Spelke et al., 2013, 2014; Walker et al., 2012). Many of the studies cited above, such as studies on *interpretive* theory of mind (e.g., Carpendale & Chandler, 1996) demonstrate that children show broad improvement in their understanding of disagreement between the ages of 5 and 8.

The ability to understand disagreements is a major part of mature scientific thinking. One's currently held beliefs might disagree with newly observed data, or two individuals might disagree about the correct interpretation for a set of observations (Barzilai & Eshet-Alkalai, 2015; Barzilai & Weinstock, 2015). Because of this, metacognition may underpin various aspects of scientific cognition (Kuhn, 1989, 2002), an argument that we outlined in chapter 1. The diagnostic reasoning measure we introduced in chapter 5 provides children with several opportunities to navigate among conflicting beliefs. For example, when shown the initial data, children might think that only the combination of all four blocks makes the machine turn green. But then they are told that there is a combination of two blocks that can make the machine turn green. Hearing this information might encourage children to change their beliefs about which blocks or combination of blocks activate the machine.

More broadly, children must learn to integrate the information they hear from others with the data they observe. In one of our investigations of this ability (McLoughlin, Finiasz, Sobel & Corriveau, 2021), we found that 5-year-olds tended to learn from an informant whose level of certainty matched the

evidence they observed (though 4-year-olds were less likely to do so). Specifically, when children observed deterministic data, the 5-year-olds we studied learned about the causal efficacy of this system better from an informant who was certain that particular events were or were not efficacious as opposed to an informant who was uncertain. When children observed probabilistic data, 5-year-olds showed the opposite pattern and learned better from informants who were uncertain than from informants who were certain. These data are part of a broader research program on children's ability to calibrate the relation between social information that indicates epistemic competence on the part of a speaker and the truth value of the speaker's claims (e.g., Birch et al., 2020; Fitneva et al., 2013; Tenney et al., 2011). All these findings show development in children's abilities to appropriately calibrate their expectations about others' testimony after the preschool years. Children develop these further metacognitive capacities to aid their scientific thinking as they enter formal schooling, and this development involves integrating together distinct kinds of data (e.g., information from verbal testimony and observed data) to reach a causal conclusion.

We wanted to investigate how developing an understanding of conflicting beliefs and how they are resolved might link to children's performance on our diagnostic reasoning measure. To do so, we presented a task investigating children's understanding of disagreement and our blicket detector task using both a cross-sectional and a longitudinal design. This allowed us to investigate the development of children's understanding of disagreement in general and to probe any possible relations between this understanding and the kind of causal reasoning necessary to solve this blicket detector task.

The Disagreement Task

In this study (Haber, Sobel & Weisberg, 2019), we wanted to investigate cases in which children had to reason about two characters who held conflicting beliefs. Specifically, we asked whether children believed that there could be legitimate disagreements, in which two characters could both have some degree of correctness, or if children take a more absolute view of beliefs. The development of this ability is interesting in its own right, but we focus on it here because it is a common situation in mature scientific thinking. Different sets of evidence might lead two researchers to draw different conclusions, or incoming evidence that does not match with one's expectations

might lead one to question or revise their previous beliefs, even if one does not reject these previous beliefs entirely.

Prior work in this area has outlined a developmental progression through different levels of understanding of disagreement, as noted in chapter 4 (Kuhn et al., 2000). Children might first believe that knowledge is just based on reality, hence beliefs must be the same for everyone (Wellman, 1990). Then they progress to an understanding that different people can have different preferences and appreciate that others' beliefs can be false, but they remain *absolutist* about other types of information: Knowledge is objective and true beliefs must be shared. On this view, when there are disagreements, one person is right and the other is wrong. Following this stage, children become *relativists* (or *multiplists*), claiming that everyone can be right about whatever belief they may hold, because all beliefs are subjective. That is, they have gained the understanding that different people can believe different things, but they still misunderstand that some of these beliefs can be more well-grounded than others. This understanding comes at the final stage of the developmental progression, called *evaluativism*, in which two people who disagree can both be right, but there is also a sense in which one can be "more right" than the other (see also Barzilai & Weinstock, 2015; Weisberg et al., 2021). The disagreement task that we used with our participants presents several situations in which characters disagree in order to probe how this developmental progression unfolds and how children's thinking about disagreements might impact their performance on the diagnostic reasoning task we described in chapter 5.

Although this progression is framed in terms of disagreements between two characters, the same idea applies to disagreements between one's own beliefs and the world or between one's current belief and a past belief. In order to think scientifically, children must understand how to reconcile disagreements between their beliefs and the data that they observe from the world, as well as between an idea that they hold currently and an idea that they held in the past (see Gopnik & Slaughter, 1991). An absolutist framework would lead them to completely accept one and completely reject the other, which is often not warranted in a true scientific investigation. Relativism is equally problematic, as there are many cases in which conflicts between different hypotheses cannot be reconciled, or in which two conflicting ideas do not have equal merit. The evaluativist framework allows children to be more successful at both understanding and conducting scientific investigations,

as they understand that older ideas can give way to newer and more correct ideas, even if the older idea still has some merit and even if the newer idea is not perfect. In this way, children's understanding of disagreements between characters can be informative as to their approach to changing their own beliefs over time.

Our measure of children's understanding of disagreement showed children two characters, attributed a belief to each character, and asked children to judge whether each character can be correct in their belief (based on Heiphetz et al., 2013; Walker et al., 2012). Using this framework, we ask about disagreements involving three different types of beliefs: beliefs about facts, beliefs about interpretations, and beliefs about preferences.

In the Fact trial, children were told about an action: An experimenter hid a penny in a certain location. They were introduced to two characters, one who said that the penny was hidden where the experimenter said that she hid it, and the other who said that the penny was hidden somewhere else. Thus, one is objectively right and the other is objectively wrong (at least if you believe that the experimenter is a reliable source of knowledge). The experimenter then asked children whether each character could be right and to justify their responses. We considered whether children responded correctly and also whether children justified their responses by referring back to what the experimenter said (*testimony-based* justifications) or by referring to the state of the world itself (*world-based* justifications).

The Interpretation trial was similar to the Fact trial, except that the experimenter's statement was ambiguous, so the penny could be hidden in one of two possible locations. One character stated that the penny was in one of the locations, while the other character stated it was in the other possible location. The penny, not being Schrödinger's cat, can be in only one of those locations. So although only one of the characters is right, we cannot know which. We again asked children whether each character could be right. We also asked them to justify their responses, which were coded into the same two categories (testimony-based and world-based) as for the Fact trial.

Finally, in the Preference trial, children were told about two characters who liked different things. Children were asked whether each character could be right about what they liked and to justify their responses. Justifications for this trial were coded as referring to the characters themselves or their preferences (*character-based*) or to the nature of preferences themselves (*preference-based*, e.g., "Because it's an opinion. There's no right or wrong").

School Partnership and Longitudinal Sample

For this line of work, we recruited students from a single school district, instead of through our labs or from museums. Working in partnership with a school district allowed us to look at how individual children's responses to our measures might change over time, as well as to address some of the issues with our previous samples. Importantly, recruiting children from museums and other venues does not allow us to control for the educational background of these participants. In contrast, for the studies described in this chapter, we know what children have been exposed to in school. We were also able to obtain these students' scores on standardized assessments in literacy, math, and science, which allowed us to determine whether there were any relations between our measures and performance on these tests.

Our school district partner for these studies was in Springfield, Pennsylvania, located in the western suburbs of Philadelphia. This district is primarily white and primarily mid- to high socioeconomic status, but it does incorporate a small degree of racial and economic diversity. In our sample specifically, 82.2% of participants were white, 6.6% were Black, 5.4% were Asian, 0.3% were Native American, and 2.0% were mixed race (3.4% of the sample did not report their race). Twelve percent qualified for free or reduced lunch.

We first began working with a cohort of 120 first graders from this district in the spring of 2014–2015 school year; we call this the 2015 group. We were able to return to the school again in the spring of 2016–2017 school year (2017 group) to retest some of these children and to recruit new participants. The 2017 sample included 78 third graders we had previously tested in 2015, providing us with a longitudinal sample. We also tested 39 third graders who had not been tested before and a new group of 112 first graders. See table 7.1 for details about these samples.

Interestingly, as part of this study, we were afforded the opportunity to link children's performance on our tasks to a change in school's curriculum. In the year between our two testing sessions (2015–2016), the school district chose to change its curriculum for the elementary grades. During our first year of testing, the curriculum was fairly traditional, emphasizing text-based content learning. This curriculum asked children to rely mainly on teachers to acquire knowledge, and students had few opportunities to actively engage in investigations, ask questions, or develop higher-level thinking or cognitive processing skills (e.g., analyzing data, constructing arguments based on

Table 7.1
Participants from the Springfield School District

	2014–2015 school year	2016–2017 school year
First graders	$n = 120$ 60 female, 60 male mean age = 86.9 months age range = 73.7 – 97.4 months	$n = 112$ 58 female, 54 male mean age = 87.3 months age range = 79.9 – 103.9 months
Third graders		$n = 117$ (includes 78 children previously tested in 2015) 59 female, 58 male mean age = 111.0 months age range = 100.6 – 120.5 months

evidence, reflecting on knowledge). Furthermore, little time was dedicated to learning social studies or science.

Over the 2015–2016 year, the school chose to incorporate more inquiry-based learning, which emphasizes the idea that children learn and actively construct knowledge through exploration, question-asking, and experimentation (e.g., Edson, 2013). This new curriculum integrated content knowledge with process skills and focused heavily on bringing more science content into the classroom, especially in the primary grades. It was structured around essential questions that focused on fostering students' critical thinking, reasoning, and analytical skills. These essential questions were designed to be open-ended, thought-provoking questions that require high-order thinking (e.g., making predictions, analyzing findings, reflecting on knowledge) in order to lead students to ask additional questions (McTighe & Wiggins, 2013; see Haber et al., 2019, for more details about the differences between these curricula as implemented in this school district).

The structure of this sample allows us to make two key comparisons. First, comparing the children who were tested both in 2015 and in 2017 gives us a longitudinal view on the development of children's understanding of disagreements and on the development of children's diagnostic reasoning. Second, comparing the first graders tested in 2015 to the first graders tested in 2017 potentially gives us insight into whether different curricula can affect children's understanding of disagreements.

Although students exposed to either a direct-instruction curriculum or an inquiry-based curriculum can learn the same content, the process by which this knowledge is acquired looks different. Instead of solely relying on the

teacher for information, children who receive an inquiry-based curriculum are obtaining their own information through direct interactions with the world. They are also asked to evaluate evidence, develop arguments, and reflect on their own knowledge as they prepare to explain their decisions to their classmates. This process of acquiring knowledge, in contrast with direct instruction, places students at the center of their own learning process. Children's experiences with these different methods of learning may thus affect their understanding of the objectivity of knowledge and hence of the ways in which different individuals may disagree.

At both time points (2015 and 2017), all participants received three tasks: the test of understanding disagreements, described above, the diagnostic reasoning measure that we introduced in chapter 5, and an interview about the meaning of the word "science." Motivation for and results from the science interview are discussed in chapter 8. In the remainder of this chapter, we focus on charting children's developing understanding of disagreements (as a proxy for metacognitive development) and their developing causal reasoning skills (as measured with the blicket detector task), and in testing relations between these two tasks.

Performance on the Disagreement Task

Figure 7.1 shows how children from all three groups responded to the three types of trials in the disagreement task. For the Fact trial, we scored responses as correct if children said that the character who agreed with the experimenter's assertion was right and that the other character was wrong. For the Interpretation trial, we scored children's responses as correct if they said that both characters could be right. For the Preference trial, we scored children's responses as correct if they said either that both characters could be right or that both characters could be wrong in having their different preferences; we only considered a response incorrect here if children said that one character was right while the other was wrong.

An in-depth discussion of the difference between the 2015 and 2017 first graders is presented in Haber et al. (2019). Here, we highlight the main finding from this comparison, which is the difference between performance on the Fact trial in these two groups. The 2017 first graders, who received the more inquiry-based curriculum, had a better understanding of situations of disagreement about a known, objective truth than the 2015 first graders.

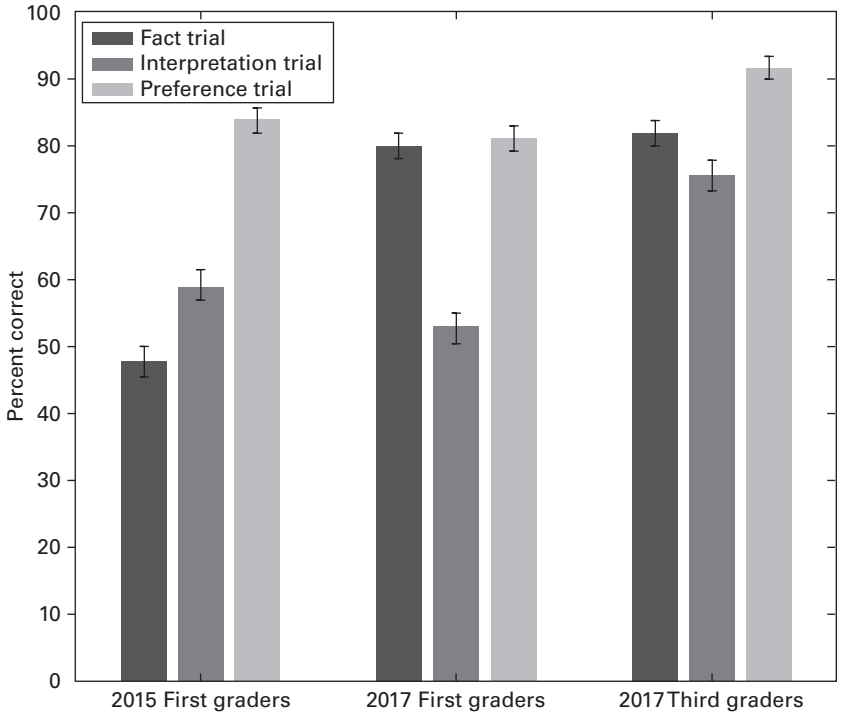


Figure 7.1
Performance of children from the Springfield School District on the disagreement task, by grade and year tested.

Further, children in these two cohorts justified their responses differently. While the proportion of testimony-based justifications did not differ between the first graders tested in 2015 and 2017, the proportion of world-based justification did. First graders tested in 2017 appealed to the actual state of the world significantly more often (39% of the time) than those tested in 2015 (16% of the time). And even though their correct responding on the Interpretation trial did not differ, a similar difference was found in their world-based justifications for this trial (32% in 2017 vs. 24% in 2015). This suggests that the curriculum difference might have focused children more on the objective truth of the situation.

We also want to discuss the contrast between the first graders and the third graders tested in 2017 (i.e., data not published in Haber et al., 2019). We found that the third graders' performance on all three trial types was

above chance.[1] Comparing these children to the first graders tested in the same year shows that third graders were better only on the Interpretation and Preference trials.[2] On the Fact trial, the first graders tested in 2017 performed in the same way as the third graders.[3] This provides evidence for the positive effect of the inquiry-based curriculum on first graders' understanding of fact-based disagreements; their performance as first graders was no difference than the performance of children who were two years older.

The pattern of justifications for the Fact trial for the third graders also did not differ from the pattern generated by the first graders tested the same year: 37% of the third graders' justifications were testimony-based and 39% were world-based (compared to 40% and 32% for the first graders).[4] However, for the Interpretation trial, the third graders generated significantly more testimony-based justifications (19%) than the first graders (8%). The third graders did not provide more world-based justifications for Interpretation trial (32%) than the first graders (32%).[5] For the Preference trial, first graders were more likely to refer to the characters and their likes and dislikes (56% of first graders, contrasted with 31% of third graders), while third graders were more likely to refer to the idea that opinions cannot be right or wrong (17% of first graders, contrasted with 59% of third graders).[6]

Longitudinal Analyses

We also considered the performance of the 78 children who participated in this task both in first grade and in third grade. The patterns we observed here generally aligned with the patterns in the data as a whole: Children improved significantly in their performance between first and third grade on the Fact and Interpretation trials, but improved only marginally in their performance on the Preference trial, likely because performance was already quite high on this trial in first grade.[7]

In terms of justifications, on the Fact trial, children in this longitudinal sample gave significantly more world-based justifications in third grade than in first grade.[8] This aligns with the findings from the larger data set, in which we also saw a relation between a greater proportion of correct responses being associated with more references to the state of the world. There were no developmental differences in children's justifications for the Interpretation trial.[9] For the Preference trial, children were more likely to justify their responses with reference to the subjective nature of opinions in general and

less likely to refer to an individual character's preferences in third grade than in first grade.[10]

Summary

The overall goal of this task was to assess how children think about disagreement across a range of different types of situations. One of the main lessons to draw from this work is the overall developmental pattern. Even first graders understand that two people can hold differing beliefs with respect to preferences. This replicates several earlier results in this area, which also show that children respect the subjective nature of preferences by the age of 7 or so (e.g., Heiphetz et al., 2014). Although the pattern of correct responding stayed relatively stable over these two years, we did find developmental changes in how children tended to justify their responses to this question. Younger children were generally more likely to think about the immediate situation, saying that a character could be right because of his or her individual preferences (e.g., "she really likes yellow"). In contrast, older children were more likely to think about the more abstract issue of how preferences work, saying that either character could be right because opinions cannot be right or wrong. This indicates that, with age, children develop a more abstract and nuanced understanding of the ways in which preferences function.

While understanding differences in two characters' preferences seems to be a relatively straightforward development achievement, understanding how to coordinate cases of conflicting beliefs about facts or interpretations is more difficult. We did see maturation here, with third graders tending to perform better than first graders. Within this general developmental context, though, the educational setting can make a significant difference. Students who were used to learning through more direct instruction were more likely to refer to the experimenter's testimony as a reason for why a character could be right. That is, the fact that these children had received a curriculum that emphasizes learning through direct testimony possibly led them to be more likely to rely on the experimenter's testimony as a source of information. A curriculum that encouraged children to engage in discovery on their own behalf seems to have made them more prone to examining the state of the world for themselves. Although there are other possible reasons for this difference, because both curricula were administered by the

same teachers in the same district, we feel reasonably confident in concluding that the curriculum played a role in this difference.

Finally, disagreements that were open to different interpretations were more difficult for children to understand than disagreements about objective matters of fact, and more difficult for children to understand than disagreements about preferences. This makes sense in light of other work on interpretive theory of mind, finding that children throughout the early elementary years have trouble understanding that two people can interpret an ambiguous situation differently (e.g., a duck-rabbit figure; Carpendale & Chandler, 1996; Mitroff et al., 2006). The ability to coordinate multiple interpretations thus may be a later developmental achievement. Indeed, this is a domain in which even adults struggle (e.g., Barzilai & Weinstock, 2015), and where development is ongoing.

Performance on Causal Reasoning Tasks

In addition to the disagreement task, we also presented some of these children with the diagnostic reasoning measure introduced in chapter 5 as a test of their causal reasoning abilities. As a reminder, in this task, children are presented with a blicket detector that lights up red or green.[11] Children observe what happens when the experimenter places different combinations of blocks on the machine. They are then asked to figure out which pair of blocks (out of three possibilities) make the machine turn green and play music, based on the data that they observed. We also asked children to justify their choices.

First Graders Tested in 2015

Half of the first graders we recruited in 2015 ($n = 60$) were given this task. We found that 45% of them provided the correct answer to this task, which is marginally better than chance performance (33%).[12] This is generally in line with the findings from our other samples (described in chapters 5 and 6), suggesting that the performance of children in those samples was not due to where or how they were recruited.

The distribution of these children's justifications also looked similar to those from the other samples. About 18% of these children made some mention of the data that they had observed. But referring to the data from the demonstration phase did not relate to their performance; about half

of the children who gave data-based justifications passed the task (55%), while children who did not give such justifications were not more successful (45%).[13] This pattern does not match what we found earlier, where data-based justifications were associated (though not significantly) with better performance. However, this sample includes only first graders, whereas the sample we discussed in chapter 5 spanned a much wider age range. It might be that these links between performance and explicit ability to justify one's own choices only appear later in development.

The other half of our first graders in this year ($n = 60$) received a different scientific thinking task: the mouse task (based on Sodian et al., 1991, which we described in chapter 4). We chose to vary which task children received because we wanted to be able to compare the performance of our participants to a published task that was already established as being a test of some aspect of scientific thinking, specifically of distinguishing hypotheses from evidence.

Like the blicket detector measure of diagnostic reasoning, the mouse task assesses children's abilities to reason backward from effects to causes. But rather than using a machine, it uses a short story that presents a reasoning problem. Specifically, children are told about two brothers who both know there is a mouse in their house. The characters want to find out whether the mouse is big or small. To do this, they make two different mouse houses. One house has a small door that only a small mouse could fit through. The second house has a large door that both a small and large mouse could fit through. The key test question asks which house the characters should choose to find out the size of the mouse. We were interested in whether children would understand that the brothers have to use the box with the small door to answer this question. Only the small mouse can get into this box, so only this box could help them to tell whether the mouse is big or small. This task assesses whether children can recognize what actions they would need to take to construct a diagnostic test.

We found that 40% of the children correctly chose the box with the small door, no different from chance (50%).[14] First graders' performance on this task is thus comparable to their overall performance than their performance on the blicket task (45% correct), although it is difficult to make direct comparisons because the mouse task presents two answer options while the blicket task presents three, making children more likely to respond correctly on the mouse task by chance.

Further, we found that children in this task tended to provide informative justifications for their choices (e.g., "because the big one can't get into this one") about 30% of the time. Unlike for the blicket task, children's justifications related to their performance on the test question, with children who responded correctly to the test question being more likely to provide informative justifications (67%) than children who responded incorrectly (6%).[15]

Similar to Sodian et al. (1991), in addition to this main test question, we also asked which box the brothers should choose to allow the mouse to get food, regardless of its size. The correct answer to this question is the box with the large door, because both mice can get into it. We found that 88% of first graders were able to answer this question correctly, showing significantly better performance on this question than on the main test question.[16] This demonstrates that children understood the parameters of the task and the relations between the sizes of the mice and the sizes of the boxes' doors. In turn, this suggests that their worse performance on the main test question was due to them not understanding how to construct a controlled experiment in this system.

In Sodian et al.'s (1991) original report of this task, children were categorized based on how they had responded to both of these questions. Those researchers found that 55% of their first graders responded correctly to both the main test question and the checking question described in the last paragraph. Thirty-five percent of their first graders responded correctly to the question about how to feed both mice, but not to the main test question of finding out whether the mouse is big or small. When we analyze our data in the same way, we see a reversal: 33% of our participants responded correctly to both questions, while 55% of our participants correctly answered the "feed" question but not the "find out" question. This might suggest that our participants are lagging somewhat behind participants in Sodian et al. (1991) in their understanding of this aspect of diagnostic reasoning. However, these differences are not statistically significant, perhaps because the sample sizes are so small (20 first-graders in the Sodian et al., 1991, study and 60 in ours).

In general, then, there are some interesting differences between the blicket task (which measures children's diagnostic reasoning abilities) and the mouse task (which measures children's abilities to evaluate a diagnostic test). Specifically, the first graders we tested who received the blicket task were able to answer correctly at above-chance levels, but generally were unable to justify

their responses in an adult-like way, suggesting that their intuitive sense of the correct answer was not yet accessible to their explicit cognition. Conversely, the first graders we tested who received the mouse task had more difficulty responding correctly overall, but those who did so were generally able to justify their responses accurately. This contrast underscores our arguments about the important role that individual tasks play in our conclusions about whether children can reason scientifically. The fact that children do so for some tasks but not others, and the fact that they can think explicitly about their decisions for some tasks but not others, highlights the need to be cautious about drawing broad conclusions about the development of scientific thinking from any single task.

First Graders Tested in 2017

In 2017, we tested a new cohort of 112 first graders on the blicket task.[17] We found that 43% of these first graders provided the correct answer, significantly better than chance.[18] As noted above, the school district adopted a new curriculum between 2015 and 2017, but the proportion of correct answers from the 2017 first graders was no different than those of the 2015 first graders.[19] This suggests some stability in this measure.

In terms of their justifications for their responses, only 15% of these children referred back to the pattern of blocks that had been placed on the machine in the demonstration phase. Unlike the 2015 sample of first graders, children in the 2017 sample who gave such justifications were significantly more likely to respond correctly to the main test question (12%) than those who did not give such a justification (4%).[20] This aligns with our findings from the studies reported in chapter 5, in which an attention to the previously observed data was associated with better performance. It is also possible that the inquiry-based curriculum that these children received might have encouraged greater attention to the data that they observed about the blocks and the machine; however, this is speculative.

Third Graders Tested in 2017

Overall, 56% of the third graders we tested in 2017 provided the correct answer, significantly better than chance.[21] This level of performance aligns with our previously reported results (chapters 5 and 6) on children of this age recruited from museum settings. The performance of the 2017 third graders was also (perhaps unsurprisingly) significantly better than performance of

the 2017 first graders (43%).[22] Again, then, we have good reason to believe that our earlier results with this diagnostic reasoning measure were not due to children being recruited from or tested in a museum setting or in a lab setting; the developmental trajectory we outlined in chapters 5 and 6 seems to be fairly robust, at least based on these data.

In terms of justifications, 26% of the third graders justified their responses with reference to the data they had previously observed, significantly more than the first graders tested in the same year (15%).[23] However, we found no relation between providing this type of justification and their performance on the blicket task.[24]

As noted above, 78 of the third-grade participants had participated in our study two years earlier, when they were in first grade. However, because some of the first graders we had tested in 2015 received the mouse task, only 36 participants received the blicket task at both time points. We found that 14 of these participants responded correctly as first graders (39%) and 17 responded correctly as third graders (47%). However, there was no relation between children's performance at the two time points: 7 children responded correctly both times and 12 children responded incorrectly both times. There were 10 children whose performance improved, responding correctly in third grade but not in first grade, but there were also 7 children whose performance declined, responding correctly in first grade but not in third grade. These numbers are small, though, so we hesitate to draw strong conclusions about these trends.

To summarize, the results from these children's performance in the blicket task align nicely with our findings from other populations. First graders (mostly 6- to 7-year-olds) were marginally above chance at choosing the correct response, while third graders (mostly 8- to 9-year-olds) were significantly above chance. We thus replicate the general developmental pattern that reasoning about this kind of complex but minimally contextualized causal system emerges starting around 7 to 8 years of age.

Relations between the Causal Reasoning Task and the Disagreement Task

As noted earlier, we included the disagreement task in this study because we anticipated that it might relate to performance on the blicket task, which served as a proxy for children's scientific thinking abilities. Understanding

disagreements is a metacognitive skill that involves children's thinking about what beliefs are and how they work—crucial processes that contribute to scientific thinking in general. We thus examined relations between children's performance with the three different types of disagreement and their responses to the blicket task.

For the first graders, we found no relations between their performance on the blicket task and their performance on the disagreement task, either when considering the three trial types individually[25] or when combining them into a single scale of overall correct performance.[26] For the third graders, by contrast, we did find a relation: Children who answered correctly on the blicket task scored significantly higher on the scale of overall correct performance in the disagreement task (mean = 2.64 out of a possible 3, SD = 0.52) than children who answered incorrectly on the blicket task (mean = 2.31 out of a possible 3, SD = 0.73).[27] We thus find some support for our hypothesis that metacognitive reasoning is related to children's scientific thinking skills. Interestingly, we found this relation only in the third graders, who were the only group to consistently score above chance on the blicket task. This further suggests that there are some developmental links between metacognitive thinking skills and scientific thinking skills.

Relations to Standardized Metrics of Academic Achievement

Although we hypothesized that children's scientific thinking might depend on their developing metacognitive capacities, it is also possible that children who answered correctly on both tasks may simply be smarter or more verbally able. Luckily, we can investigate this possibility directly. Because this project involved working in partnership with the school district, we were able to obtain our participants' scores on standardized measures of literacy and math. The main test that was used in this case was the Measures of Academic Progress (MAP), an assessment that is linked to the Pennsylvania Common Core standards. This test was administered to both the first graders and the third graders in our sample.[28] We were also able to obtain scores for a different set of standardized assessments for our group of third graders. One year after their participation in our study (i.e., when they were in fourth grade), these students were tested with a state-level assessment known as the PSSA (Pennsylvania System of School Assessment), which examines performance in English language arts, math, and science and technology.[29]

Relations with the Disagreement Task

We found some relations between performance on the disagreement task and students' scores on these standardized assessments. For first graders, higher scores on both the reading assessment and the math assessment were associated with better overall performance on the disagreement task.[30] The same was true for third graders.[31]

More specifically, first graders who responded correctly on the Fact trial earned significantly higher scores on both the reading MAP[32] and the math MAP.[33] Similarly, both first and third graders who responded correctly on the Interpretation trial earned higher scores on both the reading MAP[34] and the math MAP.[35] Third graders who responded correctly on the Preference trial achieved significantly higher scores on the reading assessment.[36]

For the PSSA test, we found positive relations between all three subtests (language, math, and science) with the third graders' overall performance on the disagreement task.[37] This latter result seemed to be driven by these students' performance on the Interpretation trial, as this was always significantly associated with higher test scores, while performance on the other two trial types was more inconsistent.

Relations with the Causal Reasoning Task

Looking at the first graders' scores on the standardized assessments in relation to the blicket task revealed no difference in performance on the blicket task based on their MAP reading scores.[38] But we did find that the first graders who scored higher on the math portion of this test tended to respond *incorrectly* to the blicket task.[39] We do not have a ready explanation for this pattern, but at the least, it suggests that skill in math is not related to the kind of diagnostic thinking abilities that are necessary to solve our blicket task. For the third graders, correct responding on the blicket task was associated with higher scores on the reading assessment[40] but not on the math assessment.[41]

With respect to the PSSA, we found, as before, that students who answered correctly on the blicket task had significantly higher language scores,[42] but there was no relation with math scores.[43] The important finding is that there was a relation with science scores: Students who answered correctly on the blicket task performed significantly better on the science section of this test.[44] This suggests that something about our blicket task is tapping into an underlying process in scientific thinking, which was our original intention for developing this task.

What Do These Data Tell Us about the Relation between Causal Reasoning and Science Education?

We began our partnership with the Springfield School District primarily for reasons of experimental control. Recruiting participants from museums and other community sites did not allow us to examine the effect of schooling, as those participants came from a wide variety of schools and educational backgrounds. By testing children within a single school district, we were able to examine the effect of curriculum in more detail, as well as to control for prior science experience at school. What we found with the blicket task in this population was in line with our findings from museums: The ability to solve this diagnostic reasoning task emerges between the ages of 6 and 9, with our first graders performing roughly at chance and our third graders performing significantly above chance. We are thus narrowing in on this age range as a critical time for the development of these sets of skills.

This is also a critical time for the development of children's advanced theory of mind skills and their metacognitive understanding, as demonstrated in prior work (e.g., Heiphetz et al., 2013, 2014; Kuhn et al., 2000; Osterhaus et al., 2016). Here we tested one facet of metacognitive thinking: the ability to understand disagreements about different types of beliefs. We found that even first graders understood that two people can reasonably disagree about matters of preference, but they did not demonstrate an understanding of disagreements about matters of fact or matters of interpretation until third grade. The one exception to this pattern was the first graders who had received a more inquiry-based curriculum. Their performance on questions about fact-based disagreement mirrored that of the third graders, illustrating that experience with exploration and investigation in the classroom can help to bolster the development of this important skill.

A more general conclusion to draw from the body of work reported in this chapter is that the relation between basic causal reasoning skills and more mature scientific thinking skills is complicated. Our diagnostic reasoning measure goes some way toward bridging this gap, bringing in a more complex set of causal structures without necessarily contextualizing these structures with particular scientific content. That performance on this measure relates to performance on a standardized measure of scientific thinking in a statewide assessment suggests that this task does relate to some aspects of scientific thinking, which is promising. But it requires

much more investigation to understand the nature of the relation and to translate these basic findings into potential interventions in the classroom. One relevant finding from this work is that there is development between the ages of 7 and 8, possibly coinciding with the development of more advanced metacognitive capacities. This suggests that we should be looking at this time point to discover other developmental shifts that may underlie the further development of scientific thinking.

Our work in this area is far from over; we have barely begun to scratch the surface of the variables that define scientific thinking and how these develop. As noted at the start of part II, we intended this set of investigations using our new blicket detector task to serve as a series of case studies for how work in cognitive development can be brought into better dialogue with work on scientific thinking abilities. We have by no means exhausted the ways in which fruitful connections can be made across these areas.

Luckily, many other researchers are starting to take up the challenge of bridging this gap and have contributed additional insights into how this process works. As reviewed in chapter 4, one major area of research within this framework is examining children's abilities to engage in belief revision: the process of weighing incoming evidence against one's existing beliefs and deciding to change one's beliefs on this basis (e.g., Bonawitz et al., 2012; Macris & Sobel, 2017; Young et al., 2012). This work generally demonstrates that young children can successfully do so under some circumstances, especially when they receive direct testimony that their initial belief was incorrect, though again there is growth in the ability to use other evidence to do so starting around age 7.

A second major area of research examines whether children understand the control of variables strategy, even if they cannot produce this strategy themselves. These studies tend to show children a confounded causal system and ask them to choose which test will allow them to draw a definitive conclusion about how the system works. This work generally finds that children can choose or recognize unconfounded experiments for simple causal systems, although they may not yet be able to produce such experiments themselves (e.g., Lapidow & Walker, 2020; see also Sobel, Benton et al., 2021, described in chapter 4). Another body of research is examining how children's nascent scientific thinking skills can be nurtured in school-based settings and transferred to real-world settings (see Sandoval et al., 2014).

Moving forward from here will involve assembling these different lines of work into more unified sets of studies that can address multiple aspects of the differences between young children's and older children's reasoning abilities. It should also involve examining in more detail the gaps between the kind of implicit causal reasoning seen in young children and the kind of explicit scientific thinking expected of older children and adults. We have begun this investigation by asking children to justify their responses, which has illustrated that children's understanding of their own actions is still quite nascent. Even at ages where children (as a group) tend to answer correctly on the test question itself, their justifications for their choices do not necessarily reflect an explicit understanding of the system's causal structure. As a first step toward a fuller investigation of this issue, part III of this book begins to address children's explicit conceptions of science, learning, and other related concepts. Our goal in that body of work is to chart the development of these explicit definitions and to examine the relation between this developmental trajectory and children's abilities to engage in causal reasoning and scientific thinking.

Constructing Science

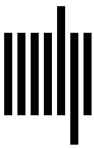
Connecting Causal Reasoning to Scientific Thinking in Young Children

By: Deena Skolnick Weisberg, David M. Sobel

Citation:

Constructing Science: Connecting Causal Reasoning to Scientific Thinking in Young Children
By: Deena Skolnick Weisberg, David M. Sobel
DOI: 10.7551/mitpress/11939.001.0001
ISBN (electronic): 9780262370615
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Weisberg, Deena Skolnick, author. | Sobel, David M., author.
Title: Constructing science : connecting causal reasoning to scientific thinking in young children / Deena Skolnick Weisberg and David M. Sobel.
Description: Cambridge, Massachusetts : The MIT Press, [2022] | Includes bibliographical references and index.
Identifiers: LCCN 2021045987 | ISBN 9780262044684 (paperback)
Subjects: LCSH: Science—Methodology. | Reasoning in children. | Scientific ability. | Science—Study and teaching—Psychological aspects. | Constructivism (Education)
Classification: LCC Q175.32.R45 W45 2022 | DDC 501—dc23/eng/20211214
LC record available at https://lccn.loc.gov/2021045987