

Children's developing theory of mind and pedagogical evidence selection

Ilona Bass¹, Alison Gopnik², Mason Hanson³, Dhaya Ramarajan, Patrick Shafto⁴, Henry Wellman⁵, & Elizabeth Bonawitz¹

1. Department of Psychology, Rutgers University, Newark, NJ 07102

2. Department of Psychology, University of California, Berkeley, CA 94720

3. Black Pine Circle School, Berkeley, CA 94710

4. Department of Math and Computer Science, Rutgers University, Newark, NJ 07102

5. Center for Human Growth and Development, University of Michigan, Ann Arbor, MI 48109

Author Note: A portion of this work was published in the Proceedings of the 39th Annual Conference of the Cognitive Science Society, and was presented as a poster at the 2017 Meeting of the Society for Philosophy and Psychology. This work was funded in part by NSF SMA SL-CN (#1640816) to EB and PS, and NSF-SES (#1627971) to EB.

Initially submitted to *Developmental Psychology* on August 17, 2017

Revised and resubmitted to *Developmental Psychology* on June 12, 2018

Revised and resubmitted to *Developmental Psychology* on September 5, 2018

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Abstract

Natural pedagogy emerges early in development, but *good* teaching requires tailoring evidence to learners' knowledge. How does the ability to reason about others' minds support early pedagogical evidence selection abilities? In three experiments ($N = 205$), we investigated preschool-aged children's ability to consider others' knowledge when selecting evidence in the service of teaching. Results from Experiment 1 revealed that four-year-olds reliably selected evidence to rectify others' false beliefs, and provided causal explanations in their teaching, whereas three-year-olds did not. In Experiment 2, we tie children's evidence selection abilities to Theory of Mind (ToM) development, above and beyond effects of age and numerical conservation abilities. In Experiment 3, we employed a 6-week training of children's pedagogical evidence selection with a new teaching task, and further explored the relationship between these skills and children's ToM abilities. We qualitatively replicated our results from Experiment 2, and report tentative evidence for a link between the pedagogical training and improvements in ToM. Together, our findings suggest important connections between reasoning about others' minds and evidential reasoning in natural pedagogy in early childhood.

Keywords: pedagogy, evidence selection, theory of mind, cognitive development, teaching, preschool children

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Children's developing theory of mind and pedagogical evidence selection

The ability to teach and learn from others is a remarkable human capability. Social transmission of information is one of the key ways in which both children and adults learn about the world; some have even argued that the natural tendency to teach, and to be ready to learn from others, may be what sets human intelligence apart from other animals (e.g., Moll & Tomasello, 2007). Indeed, humans begin to teach one another quite early in development: Three-year-olds spontaneously engage in teaching behavior with their peers (Ashley & Tomasello, 1998), and infants as young as 12 months selectively point to convey information to naïve (as opposed to knowledgeable) adults (Knudsen & Liszkowski, 2012). Investigating children's developing ability to teach others may shed insight into the cognitive mechanisms that support natural pedagogy. Here, we will suggest that the factors that support this skill – reasoning about the knowledge states of others, and reasoning about evidence – are intertwined starting in early childhood.

The early preschool years may be a prime transitional time given the development of possible components of pedagogy, such as abstract, complex, and thematic language (e.g., Gleason & Ratner, 2009; Guasti, 2002), pragmatic understanding (e.g., Barner, Brooks, & Bale, 2011; Papafragou & Musolino, 2003; Stiller, Goodman, & Frank, 2015), evidential reasoning (Bonawitz, Fischer, & Schulz, 2012; Koerber, Sodian, Thoermer, & Nett, 2005), and explicit Theory of Mind (henceforth ToM) reasoning (e.g., Wellman, Cross, & Watson, 2001). Indeed, children's teaching abilities improve considerably between the ages of three and five years. For instance, Davis-Unger and Carlson (2008) had three- to five-year-old children teach a confederate how to play a novel board game, and found that older children: 1) taught for longer

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

periods of time; 2) explained more of the rules; and 3) used a more diverse range of teaching strategies. Similarly, Strauss, Ziv, & Stein (2002) found that five-year-olds tended to teach others by providing verbal explanations, whereas three-year-olds predominantly used demonstration-based teaching strategies. There is also evidence that older children possess more declarative knowledge about pedagogy in general (Ziv & Frye, 2004). Thus, children's teaching abilities are developing rapidly during the preschool years.

This past work on children's teaching has operationalized pedagogical skills in various ways, including by the length of the teaching interaction, the types of strategies used, and whether children recognize that some individuals need to be taught while others do not. However, an additional way of conceptualizing teaching abilities comes from the distinct but related bodies of literature on concept learning, pedagogical sampling, and computational modeling. From this perspective, pedagogy is seen as a highly unique and advantageous way to learn: Instead of a learner having to sample data and learn a concept for themselves, a knowledgeable and well-intentioned teacher does the sampling for them, and then presents the learner with this selected evidence in order to help them infer the correct solution (e.g., Gweon, Tenenbaum, & Schulz, 2010; Shafto, Goodman, & Griffiths, 2014). Under this computational framework, being a good teacher requires more than just recognizing whether or not someone needs to be taught, or even that some learners need to be taught more than others; rather, good teachers must have a deep understanding of the precise evidence that certain learners may or may not need in order to infer a particular conclusion from many possible hypotheses. According to this work, then, effective teaching critically depends on selecting evidence that is tailored to the learner's prior knowledge, as well as the concept that the teacher is trying to convey (Tenenbaum & Griffiths, 2001). Indeed, this notion aligns nicely with lay theories of folk pedagogy, which

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

originally suggested that effective teachers should be sensitive to learners' knowledge and goals (Olson & Bruner, 1996).

Prior work has shown that children are sensitive to learning goals in pedagogical scenarios. Six-year-olds will select diverse samples to teach a novel concept to a peer, but not to learn that concept for themselves (Rhodes, Gelman, & Brickman, 2010). Younger children are even capable of selectively presenting evidence to intentionally deceive learners. For example, Rhodes, Bonawitz, Shafto, Chen, and Caglar (2015) showed three- to six-year-olds a novel toy that activated when any kind of block was placed on it. Children were then asked to select two blocks to demonstrate on the toy to a naïve puppet. In one condition, children were told to pick blocks that would show the puppet how the toy really worked; in the other, children were told to select blocks that would *trick* the puppet into thinking that only red blocks made it go. Across conditions, children differentially selected blocks that would best communicate their particular pedagogical goal (teach or deceive), suggesting that even very young children do consider learning goals when selecting evidence for others.

There is also work demonstrating that when learning from others, children use the evidence presented to them to make inferences about the knowledgeability of their teachers (Kushnir & Koenig, 2017; Nurmsoo & Robinson, 2009; Pasquini, Corriveau, Koenig, & Harris, 2007). Thus, as learners, children are able to monitor teachers' epistemic states in conjunction with the evidence they have presented in order to make inferences about their competence, demonstrating their abstract understanding of what it means to be an effective teacher in pedagogical scenarios. Further, an additional facet of teaching that appears to develop during early childhood is the ability to tailor instruction and provide feedback contingent on learners' particular difficulties or stage in the learning process (Strauss et al., 2002; Wood, Wood,

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Ainsworth, & O'Malley, 1995). Together, these works provide preliminary evidence that children are developing the ability to reason about tailoring evidence selections in the service of teaching during the preschool years.

However, key questions about the development of these skills are yet unanswered. Recall that from a computational perspective, “good” teaching inherently involves selecting evidence that will aid *particular* learners in arriving at specific conclusions. While there is research suggesting that children select evidence to communicate different pedagogical goals (e.g., Rhodes et al., 2015), and that children are generally sensitive to learners’ needs during pedagogical interactions (e.g., Strauss et al., 2002; Wood et al., 1995), we are not aware of any work that has asked whether children also differentially select evidence based on individual characteristics of the learner, such as prior knowledge or unobservable mental states. For example, consider the paradigm used by Rhodes et al. (2015), where children were asked to select blocks in order to communicate different pedagogical goals about a toy. Here, the learning goal (teach or deceive) was varied between subjects, and the learner’s prior knowledge was held constant, such that she was always completely naïve to the causal structure of the toy.

What if this paradigm were inverted? Now, the child’s goal would always be to communicate the true causal structure of the toy, but the learner’s *belief state* would be manipulated, such that children must select blocks to *correct* different learners’ *false* beliefs about how to activate the toy (which actually occurs when any block is placed on top of it). For instance, suppose one child sees a learner who falsely believes that only red blocks make the toy work, while another sees someone who thinks that only yellow blocks make it go. Theoretically, if these children were both “good” teachers, they would provide distinct sets of evidence to each of these two learners in order to rectify their disparate false beliefs (e.g., show the first learner

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

non-red blocks, and the second learner non-yellow blocks). However, this would require at least three highly complex inferences on the part of the teacher. First, she would need to consider that, while she understands how the toy works, the learner does not (i.e., the learner's belief is false). Second, she would need to realize that the learner *believes* he understands how the toy works, but his beliefs are incorrect (i.e., the learner thinks his belief is true). Third, she would need to select and present evidence to the learner that contradicts his particular false belief in order to lead him to the correct conclusion about how to activate the toy (i.e., understand how to select optimal evidence based on false beliefs).

Each of these three inferences requires a slightly more sophisticated level of false-belief understanding (see Wellman & Liu, 2004); the ability to reason about others' mental states (i.e., ToM) may thus be an important prerequisite for effective pedagogical evidence selection. Critically, this differs from an approach wherein the teacher merely attempts to correct the learner's behavior without considering his mental state, as learners with different belief states would require distinct sets of evidence in order to truly understand how the toy works. It could be that younger preschoolers, who are just beginning to develop explicit ToM abilities, are able to *recognize* others' false beliefs, but unable to tailor their evidence selections based on this information. Might older preschoolers be able to integrate information about learners' knowledge states into their evidence selections? What is the precise relationship between individual differences in children's patterns of pedagogical evidence selection and ToM proficiency?

To answer these questions, the current studies explore the development of preschoolers' ability to select evidence in the service of teaching across three experiments. In Experiment 1, three- and four-year-old children learned how to operate a causal toy, and then taught this causal mechanism to a confederate with a false belief about how the toy worked; we asked whether the

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

content of children's explanations provided the confederate with relevant causal information in order to correct their false belief, and also (more critically) whether children were able to consider the confederate's knowledge when selecting evidence for them (i.e., whether children selected evidence that would contradict the confederate's particular false belief). In Experiment 2, we used a similar task to investigate the relationship between children's developing ToM reasoning skills and their ability to select evidence to teach another. In Experiment 3, we further explored this relationship across a broader array of domains, and also asked whether training children's pedagogical skill might lead to improvements in ToM abilities.

Experiment 1

To examine the development of children's ability to consider others' knowledge when selecting and presenting pedagogical evidence, we designed an experiment (inspired by a task previously used by Legare, Gelman, & Wellman, 2010) in which children were presented with a novel toy that activated when red blocks were placed on platforms on top of the toy. Children were asked to demonstrate and explain to a confederate how the machine worked. In the False-Belief condition, the confederate had an incorrect belief about what types of blocks made the toy activate (e.g., square blocks instead of red blocks). In the True-Belief condition, the confederate correctly believed that red blocks activated the toy. We posit two hypotheses about children's behavior in these tasks. The first is related to the quality of children's explanations about how the toy operates, which should be considered maximally helpful when they communicate the causal rule of the toy (i.e., that red blocks make it go). Given children's general understanding that causal information is useful for learning in pedagogical contexts (Alvarez & Booth, 2015; Booth, 2009), children in our experiment who recognize that their goal is to rectify a false causal belief should be more likely to explain the causal mechanism of the toy. Because the ability to provide

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

explanations based on false beliefs does not emerge until later in the preschool years (Goodman et al., 2006), and given prior evidence that the quality of children's pedagogical explanations improves between the ages of three and five (Davis-Unger & Carlson, 2008; Strauss et al., 2002), older preschoolers should provide causal explanations more often than younger preschoolers.

Second, if children consider others' knowledge when selecting pedagogical evidence, they should show confederates with false beliefs how to operate the toy using examples that explicitly contradict that confederate's particular false belief. In contrast, when teaching a confederate with a true belief, children should not preferentially show any particular block, because any activation of the toy would be equally consistent with the learner's beliefs. A prediction that falls out of this hypothesis is that younger children, who are still developing the ability to reason about others' knowledge, may not be able to reliably select evidence that contradicts the confederates' false beliefs.

Method

Our study was approved by the University of California, Berkeley Institutional Review Board (protocol "Causal Learning in Children, 2010-01-631").

Participants. A power analysis indicated we would need twenty participants in each condition to have 80% power for detecting significant differences between groups. Therefore, sixty children were recruited from and tested at local preschools. An additional fifteen children (10 three-year-olds, 5 four-year-olds) were dropped and replaced due to: failure to pass initial memory checks ($N = 2$); failure to provide a unique final response ($N = 9$); experimental/toy error during testing ($N = 2$); or the child wanting to stop playing in the middle of the task ($N = 2$). Our final sample therefore included sixty children (20 three-year-olds: $M(SD)_{\text{age}} = 44.5(1.8)$ months, range = 41-47 months, 9 females; 40 four-year-olds: $M(SD)_{\text{age}} = 53.0(3.2)$ months, range

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

= 48-60 months, 19 females). Participants were primarily from middle-class backgrounds; a range of ethnicities resembling the diversity of the population was represented. All twenty of the three-year-olds and twenty of the four-year-olds were assigned to the False-Belief condition. The remaining twenty four-year-olds were assigned to our control True-Belief condition.¹

Materials. The causal toy was a 15.5" x 12" x 10" box covered in blue felt. Two white plastic "activator platforms", as well as two transparent spheres that lit up and spun when activated, were attached to the top of the box. The experimenter covertly operated these spinning spheres via foot pedals connected to the toy by wires. Additional stimuli included four wooden blocks of different shapes and colors (red square, red circle, yellow triangle, purple square), and two beanbag animals. See Figure 1 for a schematic of our materials and procedure.

Procedure. The child was first trained to point and provide explanations simultaneously, which they would be asked to do later in the test trial. The experimenter presented the beanbag animals and asked the child to point while answering questions requiring some explanation ("*Can you point and tell me what puppies do with a toy ball?*"). The four wooden blocks were then introduced, and the child was asked to identify the shape and color of each block. The experimenter then brought out the causal toy and said, "This is my big toy, some kinds of blocks make the toy work and other kinds do not. Want to see how it works?" She then demonstrated each of the blocks on both platforms of the toy, revealing that only red blocks made the machine work, regardless of shape.

After confirming that children understood the toy's causal rule, the experimenter explained that her friend (the first confederate), who had never seen the toy before, was waiting

¹ We did not run the additional control condition for three-year-olds, because pilot data suggested failure of the three-year-olds to pass the False-Belief condition would deem this unnecessary.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

outside; the child was asked to show the confederate how the toy worked using just one block. The experimenter “randomly” selected one of the two red blocks (either square or circle), and showed the child how to demonstrate to the confederate by putting the block first on one platform, then the other. The first platform side (left or right) demonstrated and the block selected (square or circle) by the experimenter were both counterbalanced. The experimenter then called in the confederate, who took the experimenter’s seat across from the child. The child subsequently placed the block on each platform, activating the spinning spheres one at a time. After the child placed the block on the second platform, confederates in the False-Belief condition exclaimed, “Oh, I see! This toy works with [square/circle] blocks! [Square/Circle] blocks make this toy work!” thereby expressing a false belief about the correct causal rule (that red blocks make it go). In the True-Belief condition, the confederate instead exclaimed, “Oh, I see! This toy works with red blocks! Red blocks make this toy work!” demonstrating that she correctly understood the causal rule of the toy.

The confederate excused herself, and the experimenter asked the child check questions confirming she understood the establishment of false or true belief (“*What kinds of blocks did Sophie think made the machine go?*”). If the child responded incorrectly, she was given prompts to help her recall what the confederate had said (e.g., “*Did Sophie say it was red blocks, or square blocks?*”). This kind of feedback was provided until the child could reliably restate the confederate’s belief. To control for overall exposure to each block shape, the same scenario was repeated with the second confederate using the other red block; again, confederates in the False-Belief condition expressed a false belief about the shape of the block being causally related to operating the toy, while confederates in the True-Belief condition correctly understood that red blocks made the toy work.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

The experimenter then said, "I know what would be fun, let's show our friends how the machine works with two blocks!" The child was instructed to place both red blocks (square and circle) on the machine simultaneously, such that both spinning spheres activated. The confederates then re-entered the room; the "active confederate" sat behind the machine, while the other confederate sat behind the child, under the pretense of writing something in her notebook. This was to ensure that the child would teach the active confederate in particular, and not just the only confederate present in the room. For each child, only one of the two confederates played the role of the active confederate; this was counterbalanced by which block the confederate had seen (square or circle), and which confederate played the role (first or second). The child then placed both red blocks on the machine simultaneously, activating both spinning spheres. Although both blocks were red, and thus should have activated the machine, one block's activation might be considered surprising to the active confederate (e.g., the circle block) because she incorrectly believed that the rule was that (e.g.) square blocks activate the machine. The confederate looked directly at the child and asked, "Why did that happen? Can you point and explain why that happened?" and gave the child time to respond. If the child pointed to both blocks or explained without pointing, the confederate prompted by asking, "Can you point with one hand and show me?" Once the child had pointed to one event and given an explanation, or when it became clear the child would not respond, the session concluded and the child was thanked for playing.

Predictions

Explanations. Because the child's teaching goal is related to the causal mechanism of the toy in both experimental conditions, four-year-olds should generally provide explanations with more relevant causal content than three-year-olds. These results would converge with prior work on the development of children's teaching abilities (e.g., Davis-Unger & Carlson, 2008; Rhodes

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

et al., 2015; Strauss et al., 2002), and would provide additional evidence that the abilities to consciously consider relevant pedagogical goals and provide explanations based on these goals improve throughout the preschool years.

Evidence selection. In the False-Belief condition, if children are tracking the active confederate's belief *and* using this information to guide their evidence selection, they should generally draw the confederate's attention to evidence that would help rectify her belief, by pointing to evidence that conflicts with the confederate's incorrect hypothesis. We predicted that four-year-olds should point to the belief-contradicting block more often than would be predicted by chance (50%), and more often than three-year-olds, whose evidence selections should not differ from chance. However, even if four-year-olds in the False-Belief condition do reliably provide contradictory evidence, it could be that they are simply pointing to the block that would be novel to the active confederate, without necessarily understanding that this would correct their beliefs. Examining children's responses in the True-Belief condition will allow us to rule out this possibility: In the True-Belief condition, the active confederate only saw one of the two blocks (either red square or red circle, just as in the False-Belief condition), but then expressed a true belief about the causal nature of the toy. Thus, if children in the False-Belief condition point to the block that the confederate did not see more often than children in the True-Belief condition, this will lend support to the idea that children are truly using the confederate's *beliefs* to guide their evidence selection.

Results

Because the data in the current analyses are frequencies, comparisons between groups were made using Pearson chi-square tests of independence, while comparisons to chance were made using binomial tests (all two-tailed).

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Explanations. First, we examined the types of explanations children provided to the confederate about why the toy activated. Children's explanations were coded as causal if they explicitly provided the rule that caused the toy to activate (e.g., "Red stuff makes it go"). Other responses included non-causal, descriptive explanations (e.g., "Because there's platforms" or "Because it was funny") or failures to respond (e.g., "I don't know"). Details of these response types by age and experimental condition are provided in the Supplemental Materials.

We first investigated our prediction that older children would provide more causal explanations than younger children, regardless of experimental condition. We found that while 80% of four-year-olds in the False-Belief condition provided the confederate with causal explanations for why the toy activated, only 45% of three-year-olds did so; a chi-square test of independence found that these proportions differed significantly ($X^2(1, N=40) = 5.23, p = .022, \Phi = .361$; see Figure 2). Four-year-olds in the True-Belief condition also provided causal explanations at significantly greater rates than three-year-olds (85%; $X^2(1, N=40) = 7.03, p = .008, \Phi = .419$). Thus, as children develop, they begin to reason about the kinds of explanations that may be most helpful or relevant for learners based on the nature of the learning task and the current pedagogical goal.

Evidence selection. Evidence selections were coded as the first block to which the child pointed. First, we compared the performance of three-year-olds and four-year-olds in the False-Belief condition to chance (50%). We found that 75% of four-year-olds in the False-Belief condition provided contradictory evidence, which differed significantly from chance ($p = .041$). In contrast, only 50% of three-year-olds provided contradictory evidence; this proportion did not differ from chance ($p = 1.0$). While four-year-olds were not significantly more likely than three-year-olds to provide contradictory evidence in the False-belief condition ($X^2(1, N=40) = 2.67, p =$

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

.102; see Figure 2), our differential comparisons to chance lend preliminary support to our prediction that younger children may not yet be able to consider others' beliefs when selecting pedagogical evidence.

To assess whether four-year-olds in the False-Belief condition were actually tracking the active confederate's beliefs as opposed to just pointing to the block the confederate did not see, we also examined the proportion of children in the True-Belief condition who pointed to the block that the active confederate had not seen. In contrast with the 75% of four-year-olds who provided contradictory evidence in the False-Belief condition, only 45% of the four-year-olds in the True-Belief condition pointed to the block that the active confederate had not seen before; this proportion did not differ from chance ($p = .824$). We also found that the four-year-olds in the False-Belief condition provided contradictory evidence marginally more often than the four-year-olds in the True-Belief condition ($\chi^2(1, N=40) = 3.75, p = .052, \Phi = .306$; see Figure 2). While marginal, these findings support our predictions, suggesting that older children are considering others' beliefs when selecting evidence in the service of teaching.

Discussion

Experiment 1 revealed that, consistent with past work, older children's pedagogical explanations were more causally relevant than younger children's. These findings support prior work on the development of children's teaching abilities, and provide a cursory developmental timeline for the ability to adapt the content of one's explanations to the current pedagogical goal. Critically, we also found that older children were able to select evidence to rectify others' false beliefs, whereas younger children were not. Importantly, as a requirement, all of the children included in these analyses passed the check questions about the confederate's beliefs. However, our results suggest that while some three-year-olds may be able to *track* others' false beliefs and

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

understand that they may differ from their own, they may not be able to apply this knowledge during a simple pedagogical task. Only the older children were able to use the confederate's incorrect beliefs to guide their subsequent evidential demonstrations.

While our findings suggest that evidence selection abilities are developing during the preschool years, the current results are unequipped to directly address the relationship between ToM and pedagogical evidence selection for several reasons. First and foremost, we did not assess children's ToM abilities in this experiment, as a prerequisite for inclusion required some representation of false-belief understanding (i.e., passing the check questions about the confederates' beliefs). However, it could be argued that, due to the nature of the check questions, some children might have been led into providing the correct answer about the confederates' beliefs, as opposed to truly tracking them. Additionally, our statistical comparison between three- and four-year-olds' evidence selections in the False-Belief condition did not reach statistical significance. There are several potential explanations for this: It could be due to a lack of power, or our choice to use a more spontaneous, naturalistic measure (i.e., children's pointing) to quantify evidence selections. Or, critically, it could be that ToM is a stronger determinant of children's evidence selection abilities than age, and that we *would* have seen differences in evidence selection abilities had we assessed ToM understanding, instead of using age as a rough proxy. Thus, in Experiment 2, we investigated the relationship between children's ToM development and their ability to explicitly select evidence in order to correct another's false belief, while also addressing these additional concerns from Experiment 1.

Experiment 2

There are many reasons to suspect that ToM may play a critical role in children's ability to select evidence in the service of teaching. Two central notions of ToM – reasoning about

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

others' knowledge, and intentionality – are also key aspects of effective teaching. Additionally, ToM undergoes significant qualitative change between the ages of three and five (e.g., Wellman et al., 2001; Wellman & Liu, 2004), the same period during which children's pedagogical skills are developing. Indeed, prior work has found links between ToM development and general teaching skills during the preschool years (e.g., Davis-Unger & Carlson, 2008; Strauss et al., 2002; Ziv, Solomon, Strauss, & Frye, 2016). ToM may thus be a key cognitive mechanism that supports the development of children's ability to teach others. Importantly, ToM could play an especially important role in supporting pedagogical evidence selection abilities *in particular*, which require the on-line monitoring of a learner's epistemic state relative to a specific learning goal. However, we are unaware of any work that has investigated the relationship between ToM development and children's ability to effectively select pedagogical evidence to teach others – a link that, given the age-related trends from Experiment 1, might be crucial to test.

In Experiment 2, we investigated the relationship between children's ToM development (as measured by a shortened false-belief battery; Wellman & Liu, 2004) and the ability to explicitly select evidence to correct another's false belief. Assuming this link, we predicted that children with more proficient ToM abilities would be better at pedagogical evidence selection. We used a similar paradigm as in Experiment 1, wherein children were asked to select evidence for a confederate with either a false or true belief about a toy's causal mechanism. However, instead of using children's spontaneous pointing behavior as a proxy for evidence selection, we explicitly asked children to select a single block to show the confederate on the toy.² Our hope is that this more concrete measure will allow us to more precisely assess children's understanding of how to select evidence for individuals with different belief states.

² We thank an anonymous reviewer for this suggestion.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Additionally, because we want to be able to tie evidence selection abilities to ToM in particular, and not to age or other cognitive developmental abilities, we controlled for age and also assessed children's understanding of numerical conservation (i.e., the idea that quantities remain the same despite adjustments of apparent size). Success on numerical conservation tasks, like ToM, undergoes a period of rapid development during the preschool years (e.g., Piaget, 1952), and may require maintaining dual representations (here, the quantity of an entity before and after a physical manipulation). The task may also capture other aspects of executive function such as working memory and inhibition (Miyake et al., 2000). However, numerical conservation represents a domain of cognitive development that is unrelated to social inference or reasoning about others' minds. Indeed, numerical conservation has been used as a control measure in past ToM training studies (e.g., Slaughter & Gopnik, 1996); we therefore chose to administer a numerical conservation task in the current experiment, providing support that any differences in evidence selection abilities we might see across children were due specifically to ToM abilities and not to other unrelated domains of cognitive development.

As noted above, we used false-belief tasks to measure ToM. Between the ages of three and five, children become less likely to explicitly predict others' actions based on the veridical state of the world, and more likely to understand that others' actions are in fact guided by their (sometimes false) beliefs (Goodman et al., 2006; Wellman et al., 2001). Some have argued that implicit false-belief understanding emerges at much earlier ages (between 10 and 15 months), and that apparent developments in ToM between the ages of 3 and 5 years are actually reflections of task demands (Baillargeon, Scott, & He, 2010). Nevertheless, there is evidence that the changes that occur in children's explicit ToM understanding during the preschool years are critical: During this time, children gain the ability to provide explicit causal explanations for

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

others' actions based on epistemic states (e.g., Bartsch & Wellman, 1989); further, differences in preschoolers' false-belief understanding are predictive of numerous other capabilities, including children's tendency to talk about people in everyday conversation, and their social competence more broadly (Astington & Jenkins, 1995; Imuta, Henry, Slaughter, Selcuk, & Ruffman, 2016), suggesting an important link between performance on these tasks in early childhood and real cognitive development. Therefore, we used false-belief tasks to measure ToM abilities.

Method

Our study was approved by the Rutgers University – Newark Institutional Review Board (protocol “16-625MC”). This study and subsequent analyses were also preregistered using the Open Science Framework (see Supplemental Materials for details).

Participants. We updated our power analysis using results from Experiment 1. This indicated we would need twenty-four participants in each condition to have 80% power for detecting significant differences between groups. Because we could not control whether children would pass or fail our false-belief assessment ahead of time, we collected data until we had a minimum of 24 children in each condition (but with the possibility of more children in one condition due to uncontrolled sampling). Children were recruited from and tested at local preschools. Of those tested, 16 children were dropped and replaced due to: failure to pass memory checks ($N = 10$); inability to follow the instructions of the task ($N = 1$); experimental error ($N = 3$); or the child wanting to stop playing in the middle of the task ($N = 2$). Our final sample therefore consisted of 84 children ($M(SD)_{\text{age}} = 58.9(6.1)$ months, range = 41-71 months, 41 females). Participants were from a range of socioeconomic and ethnic backgrounds that resembled the diversity of the population.

Assessments.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

False-belief. Children's ToM was assessed using three false-belief measures: a Sally-Anne task; and an unexpected contents task with questions about the self, and questions about another. In the Sally-Anne task, children were read a storybook in which a character named Sally put her teddy bear in a basket and then left the room. While Sally was gone, Alex came in and moved Sally's teddy bear from the basket to a box. Children were then asked, "When Sally comes back, where will she first go to look for her teddy bear?" Children earned a point if they correctly reported that Sally would first look in the basket. Memory checks were included to make sure children remembered where Sally initially left the bear, and where Alex moved it.

In the unexpected contents task, children were shown a crayon box, and were asked to predict what was inside. Then the experimenter revealed that rather than crayons, coins were actually inside. In the "self" portion of the task, the experimenter asked children what they thought was inside the box when they first saw it. Children earned a point if they correctly reported that they thought crayons were inside the box prior to opening it. For the "other" portion of the task, the experimenter introduced a "naïve" doll, and asked children the same question (*"When she first sees this box, what will Amy think is inside?"*). Children again received a point for correctly answering that the doll would think crayons were inside. As with the Sally-Anne task, memory check questions were included to ensure that children remembered what was actually inside the box. Thus, overall false-belief scores could range from zero to three.

Numerical conservation. To assess children's understanding of numerical conservation, experimenters showed children two parallel rows of five objects each, which were equal in length. A check question first confirmed that both rows had the same number of objects. In the subsequent two test trials, the experimenter expanded one row and contracted the other, and asked children which row had more objects, or if they were both the same. This entire process

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

was performed for two types of objects (flowers and rings), and children earned a point for every test trial in which they correctly said that the two rows had the same number of objects. Because there were a total of four test trials, conservation scores could range from zero to four.

Procedure. Children's ToM abilities were assessed at the beginning of the task. Those who scored zero or one out of three points on the false-belief battery were considered to have failed, while children who scored two or three points were considered to have passed. Thus, children were first classified as false-belief "passers" or "failers". Passers were randomly assigned to either the False-Belief condition ($N = 24$; $M(SD)_{\text{age}} = 59.4(7.0)$ months, range = 41-71 months; 14 females) or the True-Belief condition ($N = 24$; $M(SD)_{\text{age}} = 59.3(5.8)$ months, range = 46-71 months; 8 female), whereas failers ($N = 36$; $M(SD)_{\text{age}} = 58.3(5.8)$ months, range = 45-67 months; 19 females) were all run in the False-Belief condition. Following the false-belief assessment, children's numerical conservation abilities were also measured.

The remainder of the procedure was highly similar to the causal toy teaching task used in Experiment 1. Children were introduced to four wooden blocks: a red square, a red circle, a yellow triangle, and a purple square or circle (depending on counterbalancing condition; see below). Then, the experimenter introduced the causal toy, which was identical to the toy used in Experiment 1, except it had only one activator platform and spinning sphere instead of two. (See Supplemental Materials for a photograph of these materials.) The child was then asked to try all four blocks on the activator platform, revealing that only red blocks made the toy work. After ensuring that the child understood this causal rule, the experimenter explained that they were going to watch a video of her friend, who had found this toy and played with it the other day, but didn't know the rule before playing with it. The experimenter then played a short video in which the first confederate sat at a table with the toy and a single red block (either square or circle,

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

counterbalanced). The confederate tested this block on the toy, which caused the toy to activate. In the False-Belief condition, the confederate exclaimed, "Oh, I see! This toy works with [square/circle] things! [Square/Circle] things make this toy work!" In the True-Belief condition, the confederate instead inferred the correct causal rule (*"This toy works with red things!"*).

To control for overall exposure to each block shape, the experimenter also played another video of a second confederate, in which the red block on the table was the shape that had not been seen in the first video. Again, confederates in the False-Belief condition expressed a false belief about the block's shape causing the toy to work, while confederates in the True-Belief condition correctly understood that red blocks made the toy work. After watching both videos, the experimenter presented pictures of the two confederates that the child had just watched. The child was then asked to remind the experimenter of both confederates' beliefs (*"What did Jesse think the rule was about how the toy works?"*), and the true causal rule of the toy (*"What's really the rule about how the toy works?"*). Children who required more than 2 prompts for these final check questions were dropped from analysis.

The experimenter then told the child that she was going to see one of the two confederates (counterbalanced) later today, and she could bring her toy with her to show them, but could only bring one block to put on the toy to make sure the confederate understood the rule. The child was then presented with three blocks (red circle, red square, yellow triangle), and was asked to pick which block the experimenter should show the confederate on the toy. As in Experiment 1, the dependent measure was whether children selected the object that would both activate the toy and, critically, correct the confederate's false belief.

We were also interested in whether children could use a more sophisticated strategy of presenting *negative* evidence to correct another's beliefs. To test this, we added a second test

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

trial. After the child made their first selection, the experimenter replaced the “correct” red block (i.e., the block that the confederate had not seen in the video, which varied by counterbalancing condition) with the purple block that the child had seen at the beginning of the task, and asked the child to make a second selection. This purple block was the same shape as the block that the test confederate had seen in the video, and could thus hypothetically correct the test confederate’s false belief by providing contradictory *negative* evidence, showing that shape was the incorrect explanatory variable (although this would not provide positive evidence for color).

Predictions

Our predictions are akin to those from Experiment 1. First and foremost, in the False-Belief condition, passers should select the belief-contradicting block more often than would be predicted by chance (33%), and more often than failers, whose evidence selections should not differ from chance. Critically, evidence selections should *not* differ between these groups when split by age or understanding of numerical conservation. Additionally, children in the True-Belief condition should not differ significantly from chance, because these children were choosing evidence for a confederate who already correctly understood the toy’s causal rule.

Results

Following our preregistered analysis plan, comparisons between groups were made using Pearson chi-square tests of independence, while comparisons to chance were made using binomial tests. Although we preregistered our predictions, we maintained two-tailed tests here for conservatism and consistency with the rest of the manuscript.

First test trial.

Comparisons to chance. First, the proportion of children who selected the block that the test confederate had not seen was compared to chance (33%) in each of the three experimental

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

groups (failers in the False-Belief condition; passers in the False-Belief condition; passers in the True-Belief condition). Results aligned with our predictions: 58% of the 24 passers in the False-Belief condition selected the correct block, which differed significantly from chance ($p = .021$). In contrast, only 28% of the 36 failers and 46% of the 24 passers in the True-Belief condition selected the correct block; neither of these proportions significantly differed from chance ($ps \geq .280$). See Figure 3 for a summary of these proportions.

Group comparisons. Next, we asked whether the proportion of passers in the False-Belief condition who selected the correct block was significantly greater than the proportion of failers who did so (a comparison that was not statistically significant when using age to predict evidence selection abilities in Experiment 1). We found that passers selected the correct block significantly more frequently than did failers ($X^2(1, N=60) = 5.60, p = .018, \Phi = .306$), suggesting that ToM may indeed be a key cognitive factor that supports evidence selection abilities in early childhood. We also compared passers in the false-belief and True-Belief conditions to one another, but did not find significant group comparisons ($X^2(1, N=48) = .75, p = .386$). We return to potential explanations for this result below.

Effects of age and conservation. While the results thus far seem to suggest that ToM development supports the ability to pedagogically select evidence, there are other factors that could potentially be driving this effect, including age or more general cognitive abilities. To investigate these possibilities, we first found that passers and failers in the False-Belief condition did not differ from one another in terms of average age ($t(58) = .63, p = .531$) and that, using a median split by age, the proportion of older children selecting the correct block did not differ from the younger children ($X^2(1, N=60) = 2.5, p = .11$). Further, splitting data in the False-Belief condition by conservation abilities (Low = 0-1, $N = 39$; High = 2-4, $N = 21$) also did not yield

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

significant differences between groups ($X^2(1, N=60) = .11, p = .740$). Thus, above and beyond effects of age and some other cognitive factors, ToM appears to be a key predictor of children's evidence selection abilities.

Second test trial. We also administered a second test trial, in which we asked whether children could perform a more sophisticated form of evidence selection by choosing a block that would provide *negative* evidence to correct the confederate's false belief. However, only four children in total made the correct selection on this second test trial. Therefore, it seems that this more nuanced form of evidential reasoning may still be developing during the preschool years.

Discussion

In Experiment 2, we report critical evidence for a relationship between ToM development and pedagogical evidence selection abilities. In particular, we found that children who passed our false-belief battery were more likely to select evidence that would correct another's false belief than children who failed the battery. Critically, evidence selection abilities did not differ by age or conservation scores, suggesting that ToM in particular is closely tied to the ability to select evidence in the service of teaching another.

We did not find significant differences between passers in the False-Belief (58%) and True-Belief conditions (46%), although there was a nonsignificant trend. One explanation for this result is the phrasing of the test question: Children were asked to select a block to "make sure" the confederate understood the rule. Although confederates in the True-Belief condition already correctly understood the toy's causal rule, showing them the block that they hadn't previously seen would indeed "make sure" that they completely understood how the toy worked. This small pragmatic idiosyncrasy may explain why we do not see significant differences in evidence selections between the two conditions in the current experiment.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

There has been important prior work demonstrating the links between reasoning about others' minds and pedagogy (e.g., Strauss et al., 2002; Wood et al., 1995); here, we provide preliminary evidence in establishing the relationship between pedagogical evidence selection in particular, and ToM development. However, critical questions about the nature of this relationship remain. For instance, Experiments 1 and 2 both asked children to select evidence to correct another's false belief, in the context of a causal toy activation task. What about evidence selections that do *not* entail false belief correction, or that utilize tasks in different domains? Currently, we cannot know whether ToM would predict evidence selection abilities outside of the particular set of circumstances tested in our first two experiments. Additionally, our measures of evidence selection have thus far been binary (e.g., whether the child selected the correct block, or not). However, pedagogical skill as it pertains to evidential selection is a highly nuanced construct; further, these binary measures did not allow us to statistically control for age and conservation abilities as a continuous variable in the current analysis. In order to assess more subtle developments of these skills, different measures are needed. Finally, we have been largely assuming that it is ToM development that supports pedagogical skill, but have not explicitly tested the direction of this relationship. It is possible that these abilities develop together because they mutually influence each other. For example, practice with evidential reasoning could potentially drive the development of ToM by exposing children to the idea that the learner's belief state is dependent on the evidence observed (which may not reflect "truth"). In Experiment 3, we employed a 6-week training of pedagogical evidence selection using two different tasks and a more graded measure of pedagogical skill, in order to explore these open questions.

Experiment 3

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

There is some existing evidence to hint that training pedagogical skill might lead to improvements in ToM abilities. In particular, asking children to produce explanations for characters' behavior in false-belief tasks promotes the development of false-belief understanding (Amsterlaw & Wellman, 2006; Wellman & Lagattuta, 2004). Although explanatory reasoning may improve ToM for myriad reasons, including deepening causal commitments (e.g., Walker, Lombrozo, Legare, & Gopnik, 2014), given that providing explanations is itself a type of pedagogical strategy, we might expect to see children's false-belief understanding improve as a result of pedagogical training. However, it remains an open question as to how evidential reasoning in the service of teaching may be related to ToM.

The current study extends this prior finding on false-belief development to distinct domains (word learning and causal reasoning), asking whether training pedagogical evidence selection leads to improvements in children's understanding of false-belief. To investigate this, we assessed children's false-belief understanding, and compared their performance on two teaching tasks. Children who did not pass the false-belief tasks were subsequently trained on these teaching tasks for six weeks, after which their false-belief understanding was reassessed. As in Experiment 2, we controlled for age and also assessed children's understanding of numerical conservation.

Method

Our study was approved by the University of California, Berkeley Institutional Review Board (protocol "Causal Learning in Children, 2010-01-631").

Participants. A power analysis indicated we would need twenty participants in each condition to have 80% power for detecting significant differences between groups. Therefore, sixty-one children were recruited from and tested at local preschools. An additional thirteen

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

children were dropped and replaced due to: attrition during the six-week training ($N = 3$); failure to pass memory checks ($N = 1$); experimental error ($N = 1$); or the child wanting to stop playing in the middle of the task ($N = 8$). Our final sample therefore consisted of sixty-one children ($M(SD)_{\text{age}} = 47.1(4.8)$ months, range = 39-54 months, 34 females). Participants were primarily from middle-class backgrounds; a range of ethnicities resembling the diversity of the population was represented.

Tasks.

False-belief & numerical conservation. Children's ToM and numerical conservation abilities were assessed using the same measures from Experiment 2. Stimuli for these tasks varied slightly between pre- and post-test (see Supplemental Materials for details).

Pedagogical training and test. The pedagogical training entailed a novel word learning task and a causal toy activation task. In the novel word learning task, children were told that they would be teaching a confederate about a novel word (e.g., *Dax*), which represented some concept they were trying to learn. They were then shown a picture of an object with two discrete features (e.g., a *fork* that is *white*), and were told that this represented the novel word ("*Here is an example of the word Dax.*"). The word's extension was intentionally ambiguous: It was equally likely that *Dax* could mean white objects, or forks, or white forks. Given this ambiguity, the experimenter explained to the child what the novel word really meant (e.g., "*Dax means fork.*"). The experimenter then presented three additional pictures, two of which contained an item that overlapped with just one of the original picture's two features (e.g., a white spoon, and a black fork), and one of which was a distractor item that shared no features with the original picture (e.g., a black spoon). The experimenter indicated which of these four pictures did and did not

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

represent the novel word, after which she asked the child whether each picture was an example of the word, to ensure that the child understood the novel word's extension.

The experimenter then removed the distractor item, and told children they were now going to teach a confederate about the novel word using the three pictures (i.e., providing examples without explicitly telling the confederate what the novel word meant). In order to provide a correct response, children had to indicate which pictures were examples of the novel word, and which were not, thereby demonstrating all necessary and sufficient examples to identify the correct rule while ruling out other hypotheses.³ The novel words' extensions differed across trials: Some words' extensions were the conjunct concepts (e.g., white fork), while other words represented a single concept dimension (e.g., any white object, or any fork). Additionally, all of the novel words in this task represented familiar (as opposed to novel) objects. Indeed, throughout the preschool years, children often confront the task of learning new, additional names for objects that already have a prior familiar name (overcoming their early mutual exclusivity bias). Because this task required children to *teach* these new word extensions, we used categories with which children were already familiar in order to reduce the cognitive demands associated with this teaching task.

In the causal toy activation task, children were presented with a novel toy with two distinct mechanisms (e.g., a wheel and a bell). The experimenter first showed children how to activate the toy, causing it to perform some desirable outcome such as lighting up or playing music (*"You need to ring the bell and spin the wheel at the same time to make the toy go."*). As

³ To help children understand the hypotheses under consideration, the confederate announced the full set of hypotheses before the child began teaching (e.g., *"I see, Dax could mean any fork, or Dax could mean anything that is white, or Dax could mean just white forks. Can you teach me?"*).

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

in the novel word task, after children understood how to operate the toy, they were then instructed to teach a confederate about the toy by providing examples of which combinations of mechanisms did and did not make the toy go. In order to provide a correct response, children had to demonstrate both necessary and sufficient evidence for the confederate to rule out all alternative explanations and correctly infer which mechanism(s) activated the toy. As with the novel word task, some toys required the conjunction of two actions (e.g., the wheel and the bell together) while others required just a single mechanism (e.g., anytime the wheel was spun, or anytime the bell was rung).

For both tasks, if children provided insufficient evidence, the confederate prompted the child by musing aloud about the remaining possible hypotheses. For example, if the child only showed the confederate that operating both mechanisms simultaneously made the toy go, the confederate might say: "Oh, so you showed me both at the same time. It could be that you need to do both at the same time to make it go, or it could be that the wheel by itself could make it go, or that the bell by itself could make it go. Can you teach me?" Note that children would often need to present negative examples to rule out plausible hypotheses (e.g., showing that the wheel by itself did *not* activate the toy). Given that children appeared to have difficulty with negative evidence selection in Experiment 2, we expect this task to be challenging. The number of prompts children required before providing complete evidence was the primary dependent measure for both pedagogical training tasks; these scores could range from a minimum of zero (i.e., children who provided necessary and sufficient evidence spontaneously) to a maximum of two (i.e., children who required prompting after each demonstration until all evidence had been

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

provided).⁴ This measure thus provides a more graded assessment of pedagogical skill, while still capturing the degree to which appropriate evidence was spontaneously selected.

There were six different versions of each task (see Supplemental Materials for details). As described above, the different versions of each of these tasks required children to teach either a conjunct rule or a one-dimensional rule; varying the stimuli in this way ensured that children would have distinct teaching goals on different trials, and would thus have to select evidence that corresponded to the particular teaching goal of a given trial in order to provide a correct response. Note also that in the causal toy task, children did not necessarily need to exhaustively demonstrate all possible pieces of evidence on every trial in order to provide a correct response. For instance, if the child's goal was to show that the causal toy activated any time the wheel was spun, she needed only to show that the wheel activated the toy and that the bell did not (but not the result of the wheel and bell activated together), in order to rule out all remaining possible hypotheses about how the toy worked. The novel word tasks entailed exhaustive demonstrations.

Procedure. Children's understanding of false-belief and numerical conservation was assessed on a preliminary testing day. As in Experiment 2, children who scored zero or one out of three points on the false-belief battery were classified as "failers", while those who scored two or three points were considered "passers". Children who failed ($N = 40$) were randomly assigned to either the control or the training condition; the average ages of children in each of these conditions did not differ significantly ($p = .65$). Over the course of the following six weeks (beginning on the preliminary testing day), children in the training condition ($N = 22$; $M(SD)_{\text{age}} =$

⁴ Some children provided redundant evidence after receiving a prompt, and thus may have required more than 2 prompts on certain trials. These trials were coded as having required 2 prompts, making 2 the maximum for our measure of pedagogical skill.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

46.3(4.5) months, range = 39-53 months; 15 females) received two training sessions per week on both pedagogical tasks. One version of each task was administered on a given testing session, with the novel word task always being presented first. As there were six versions of both the novel word and causal toy tasks, the experimenter administered the same version of each task across both sessions of a given week. The order in which the different versions of the tasks were presented was randomized across participants. At the end of this six week period, children's understanding of false-belief and numerical conservation were reassessed.

Failers in the control condition ($N = 18$; $M(SD)_{\text{age}} = 45.6(4.3)$ months, range = 39-54 months; 7 females) received no pedagogical training, and their false-belief and conservation understanding was reassessed after a six-week delay. This ensured that the same amount of time elapsed between pre- and post-test for false-belief failers in both the training and control conditions, allowing us to control for any naturally occurring cognitive development that took place over the course of those 6 weeks. False-belief passers ($N = 21$; $M(SD)_{\text{age}} = 49.3(5.0)$ months, range = 40-55 months; 12 females) did not receive longitudinal pedagogical training, since they had little to no room for improvement on the false-belief tasks; instead, they received just one session of the pedagogical tasks on the preliminary testing day, allowing us to measure their initial teaching abilities. The versions of the pedagogical tasks used with passers were randomized across participants. See Figure 4 for a schematic of our study design.

Predictions

Given past work on the relationship between ToM and teaching abilities (Davis-Unger & Carlson, 2008; Strauss et al., 2002; Ziv et al., 2016), as well as the results from our previous two experiments, we predicted that children with more proficient false-belief understanding would also be better at pedagogical evidence selection on the preliminary testing day (i.e., prior to any

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

training). We also predicted that the pedagogical training might lead to improvements in ToM reasoning abilities; in particular, children who initially failed the false-belief battery should improve more in false-belief understanding if they took part in the pedagogical training than if they experienced a 6-week delay.

Results

One failer in the training condition did not complete one session of the causal toy task, another failer did not complete one session of both tasks, and one passer's numerical conservation abilities were erroneously not recorded. These individual data points were treated as missing in subsequent analyses. Otherwise, all children completed all training sessions and assessments.

We created a composite pedagogical skill score for each training session by calculating the average number of prompts children required across both tasks in a given session. Lower scores indicated better task performance (i.e., more advanced pedagogical skill). With the exception of the classification of children as false-belief passers or failers, all measures were treated as continuous in the following analyses (all tests two-tailed).

Initial false-belief understanding & pedagogical skill. We first investigated the effects of preliminary false-belief understanding on initial (i.e., non-trained) pedagogical skill, as measured by the average number of prompts children required across both tasks on the first testing session. An independent-samples t-test compared the average number of prompts required on the preliminary testing day between false-belief failers in the training condition and false-belief passers. As with Experiment 2, we found a relationship between false-belief score and evidential reasoning proficiency: Passers ($M = 1.05$, $SD = .57$) provided necessary and sufficient evidence with significantly fewer prompts than failers ($M = 1.45$, $SD = .55$), $t(41) = 2.38$, $p =$

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

.022, Cohen's $d = .725$; see Figure 5. Looking at the novel word and causal toy tasks separately, we found that this effect was driven by differences in performance on the causal toy task (Passers: $M = .57$, $SD = .68$; Failers: $M = 1.27$, $SD = .77$), $t(41) = 3.17$, $p = .003$, Cohen's $d = .968$. A chi-square test of independence further revealed that, on the causal toy task, false-belief passers required zero prompts (i.e., spontaneously provided necessary and sufficient evidence) more often than failers, while failers were more likely to require at least two prompts before providing complete evidence ($\chi^2(2, N=43) = 8.58$, $p = .014$, $\Phi = .447$; see Figure 6).

In contrast, an independent-samples t-test revealed no significant differences between passers and failers on the novel word task (Passers: $M = 1.52$, $SD = .75$; Failers: $M = 1.64$, $SD = .58$, $p = .584$). This disparity may be explained by the increased difficulty of the novel word task relative to the causal toy task. Indeed, two paired-samples t-tests revealed that both passers and failers performed better on the causal toy task than on the novel word task on the preliminary testing day (Passers: $t(20) = 2.16$, $p = .042$; Failers: $t(21) = 5.05$, $p < .001$); additionally, more children required at least two prompts on the novel word task ($N = 29$) than on the causal toy task ($N = 12$). We return to possible explanations for the difference in difficulty across these two tasks in the General Discussion.

Supporting our results from Experiment 2, false-belief passers generally outperformed failers in terms of their pedagogical skill. However, there are many possible reasons why passers may have outperformed failers on the causal toy task, including age or other cognitive factors. To control for this, we ran two between-subjects ANCOVAs, with false belief proficiency (passers vs. failers) predicting performance on the causal toy task; we included preliminary conservation scores as a covariate in one analysis, and age at pre-test in the other. Passers still outperformed failers on the causal toy task, even when controlling for effects of age ($F(1, 40) =$

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

6.11, $p = .018$, $\eta_p^2 = .133$) and conservation scores ($F(1, 39) = 9.35$, $p = .004$, $\eta_p^2 = .193$). These results provide even stronger evidence for a direct link between false-belief understanding and teaching abilities.

Effect of training on pedagogical skill. Next, we began to investigate the different potential effects of the pedagogical training. Preliminarily, we were interested in whether the training was actually effective in improving children's pedagogical evidence selections. Using data from false-belief failers in the training condition, a repeated-measures ANOVA on the mean number of prompts required on each of the twelve training sessions revealed a significant effect of session ($F(11, 220) = 4.96$, $p < .001$, $\eta_p^2 = .20$), as well as a significant linear trend (i.e., a straight line fit the data at better than chance levels; $F(1, 20) = 20.85$, $p < .001$, $\eta_p^2 = .51$; see Figure 7). Additionally, a paired-samples t-test found that children required significantly more prompts on the first day of training than on the last day ($M(SD)_{\text{FirstDay}} = 1.45(.55)$; $M(SD)_{\text{LastDay}} = .64(.49)$, $t(21) = 5.36$, $p < .001$). We also looked at the causal toy and novel word tasks separately. Two repeated-measures ANOVAs revealed significant effects of session for both tasks (Causal toy: $F(11, 209) = 2.26$, $p = .013$, $\eta_p^2 = .11$; Novel word: $F(11, 220) = 4.35$, $p < .001$, $\eta_p^2 = .18$), and two paired-samples t-tests also found that children required fewer prompts on the last day than on the first day for both the causal toy task ($M(SD)_{\text{FirstDay}} = 1.27(.77)$, $M(SD)_{\text{LastDay}} = .59(.67)$, $t(21) = 3.38$, $p = .003$) and the novel word task ($M(SD)_{\text{FirstDay}} = 1.64(.58)$, $M(SD)_{\text{LastDay}} = .68(.65)$, $t(21) = 5.31$, $p < .001$). Children's performance on the pedagogical tasks thus did improve with training, suggesting that these tasks represent a viable method for improving children's ability to select evidence in the service of teaching.

Relationship between false-belief improvement & aggregate pedagogical skill. Next, we investigated the relationship between overall aggregate performance on the two pedagogical

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

tasks and improvement on the false-belief battery from pre- to post-test. Using data from false-belief failers in the training condition, we ran a correlation between false-belief improvement (i.e., pre-test false-belief scores subtracted from post-test scores) and the mean number of prompts required across all twelve training sessions. We found a statistically significant negative linear relationship between these two factors ($r(20) = -.43, p = .047, R^2 = .182$; see Figure 8). In other words, children who required fewer prompts over the course of the training generally improved more in false-belief understanding from pre- to post-test. Two partial correlations revealed that this finding qualitatively persisted when statistically controlling for average age ($r(19) = -.43, p = .054, R^2 = .182$) and improvement in conservation understanding ($r(19) = -.41, p = .063, R^2 = .171$).

Effect of training on false-belief understanding. Finally, we evaluated whether the pedagogical training was successful in improving children's false-belief understanding. We ran an independent-samples t-test comparing false-belief failers in the training condition ($N = 22$) to those in the control condition ($N = 18$) on false-belief improvement (i.e., pre-test false-belief scores subtracted from post-test scores). This direct comparison between training and control participants did not yield significant results: Children in the training condition ($M_{\text{Improve}} = .41, SD = 1.0$) did not improve in false-belief understanding any more than did children in the control condition ($M_{\text{Improve}} = .56, SD = .98, p = .65$). In fact, children in *both* conditions improved at least marginally in false-belief understanding from pre- to post-test (Training: $t(22) = 1.90, p = .071$, Cohen's $d = .406$; Control: $t(17) = 2.40, p = .028$, Cohen's $d = .565$).

Because our training included both children who failed all false-belief tasks initially and those that passed one, we hypothesized that partial passers may have been at a more "liminal," or transitional, stage of ToM development (Goodman et al., 2006), and thus might have been more

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

likely to improve in false-belief understanding regardless of experimental condition. This improvement across conditions due to maturation may have effectively “washed out” the possible effects of the training. Thus, we constrained our analysis to children who answered zero false-belief questions correctly at pre-test. Results revealed improved false-belief understanding for children in the training condition ($N = 12$; $M_{\text{Improve}} = .58$, $SD = .79$; $t(11) = 2.55$, $p = .027$, Cohen's $d = .735$), whereas no significant improvement was seen for children in the control condition ($N = 8$; $M_{\text{Improve}} = .25$, $SD = .46$; $p = .170$); see Figure 9. Importantly, conservation scores did not differ for either group between pre- and post-test (Training: $p = .551$; Control: $p = .197$), suggesting that the pedagogical training targeted ToM without necessarily leading to general improvement in cognitive reasoning. Note that this result does not directly compare training to control children, and must therefore be interpreted with caution. However, coupled with our finding that initial false-belief understanding is related to non-trained pedagogical skill, as well as our results from Experiment 2, this may suggest a critical (and potentially bidirectional) link between reasoning about others' minds and pedagogical evidence selection in early childhood.

Discussion

Our results from Experiment 3 suggest that having a more developed ToM is broadly related to being better at pedagogical evidence selection – even in the context of teaching goals that do not entail correcting another's false belief. This finding persisted when statistically controlling for age and understanding of numerical conservation. Further, we found tentative supporting evidence for the idea that training pedagogical evidence selection may in turn improve children's ToM reasoning abilities. Importantly, it is unlikely that our pattern of findings could have been achieved without children having some representation of the learner's

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

knowledge state. For instance, consider the explanation that children were simply demonstrating their own knowledge, as opposed to “truly teaching.” If this were the case, we might expect that upon receiving a prompt from the confederate, children would simply re-demonstrate what they had just shown. However, our data show that at a minimum 72% of participants in the training condition provided novel evidence after each prompt in the causal toy task (see Figure 6). Taken together, our results support and extend our findings from Experiments 1 and 2, converge with prior work on the relationship between ToM and teaching skills, and suggest that there is indeed a strong developmental link between pedagogical evidence selection and ToM.

General Discussion

Past work has shown that children's developing ToM reasoning abilities are related to their pedagogical skill, but has not looked at the precise relationship between the development of children's capacity to reason about others' minds and the ability to select optimal evidence to teach others. We investigated this link across three experiments. In Experiment 1, we found that only older preschoolers provided evidence and explanations that would correct their learner's particular false belief. Experiment 2 tied evidence selection abilities directly to ToM development, which predicted evidence selections more reliably than age or numerical conservation abilities. In Experiment 3, we found that children with more proficient ToM reasoning abilities were also better evidence selectors in paradigms that did not entail false-belief correction, and using a more graded measure of pedagogical skill. Additionally, training evidence selection skills led to significant increases in ToM reasoning abilities for children who had the most room for improvement in ToM at pre-test. Results from our experiments suggest that there may be important and deep connections between reasoning about others' minds, evidential reasoning, and natural pedagogy in early childhood.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Our findings have clear implications for current theories and models of natural pedagogy and epistemic trust. Shafto et al. (2014; Shafto & Goodman, 2008) propose a Bayesian model of pedagogical teaching and learning, according to which the evidence that teachers choose to present directly depends on the learner's prior knowledge and the learning goal that the teacher is trying to communicate. This pedagogical model is a special case of the broader model of epistemic trust (Eaves & Shafto, 2012, 2017; Shafto, Eaves, Navarro, & Perfors, 2012), which explicitly connects developmental changes in reasoning about others' beliefs to interpretation of evidence selection by others. Our results support these models that link evidence selection and reasoning about others' minds. We also extend their findings, showing that this link: 1) exists even in young children who have not yet been exposed to formal schooling; and 2) is manifest in their selection of evidence for others.

Our results also speak to existing models of ToM development that postulate genuine conceptual change during the preschool years. Specifically, in Experiments 2 and 3, we found evidence for a link between performance on a false-belief task and the discrete developmental capability of pedagogical evidence selection, suggesting that the changes in false-belief understanding that occur between the ages of 3 and 5 may reflect deep qualitative changes in children's ToM. Of course, we recognize that there is a diverse range of perspectives on the course of children's ToM development, and we will not attempt to resolve that debate here. Rather, we simply suggest that our findings cannot be explained entirely by false-belief task demands (especially given that our results persist when controlling for effects of age), and may therefore be indicative of some type of conceptual change in ToM between the ages of three and five years.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

In Experiment 2, we added a second test trial to investigate whether children understood that negative evidence could also correct another's false belief, but found that almost no children made this selection. Interestingly, however, the pedagogical tasks in Experiment 3 often *required* children to provide negative evidence, in order for the confederate to be able to infer the correct hypothesis. While young children generally rely heavily on negative evidence for making inductive inferences themselves (Kalish & Lawson, 2007), our results suggest that the ability to *provide* negative evidence may be contingent upon context, and could still be developing during the preschool years.

An additional unanticipated finding from Experiment 3 was that while children did improve on both pedagogical tasks over the course of the training, they did not perform equivalently across the two tasks on the preliminary testing day; specifically, while false-belief passers outperformed failers on the causal toy task, children generally performed worse on the novel word task, regardless of ToM proficiency. While this disparity could be attributable to the stricter criteria for "success" in the novel toy task, there are also several possible deeper explanations. First, there could be some underlying difference in the nature of the tasks that made the novel word task more difficult for young children. For instance, while it was technically possible for children to present evidence without any verbal communication in the causal toy task, the novel word task might have been slightly less conducive to that strategy. Given the natural emergence of verbal explanations in pedagogy during the preschool years (Strauss et al., 2002), it could be that the verbal nature of the novel word task made it excessively difficult for the younger children in our study.

Another possible explanation for the disparities between these two tasks is the relative abstractness of the novel word task, in which children were required to communicate the

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

extension of an intangible concept. Given that abstract reasoning is still rapidly developing during the preschool years (e.g., Holyoak, Junn, & Billman, 1984), this could be another reason that the novel word task proved to be difficult for the children in our sample. A final possibility is that preschool-aged children have more experience with teaching causal effects by demonstration (of the type demanded by our method) than with teaching word meanings by demonstrations of positive and negative extensions. Word learning in the preschool years may often be given by a teacher explicitly providing a linguistic synonym (e.g., “A spork is like a fork and a spoon”) and rarely by exemplars alone (e.g., “This is an example of a spork, that is not an example of a spork”). On the other hand, causal learning, such as when exploring a new toy, may comparatively more often unfold via demonstration without accompanying linguistic specification (e.g., “Try this to see what it does!”). Thus, it is possible that children’s past experience with learning from others in these two domains provided a more familiar context for causal teaching from demonstration than word learning from demonstration alone. Future work might consider teasing apart these possible explanations, as this could shed insight into the development of children’s ability to select different kinds of evidence (e.g., abstract vs. concrete; verbal vs. non-verbal; etc.).

While we used a numerical conservation task as a control for more general cognitive development, there are other possible aspects of development we could have assessed. Although we designed a task that required minimal linguistic demands on our participants, language proficiency scores might provide additional insight into our results, especially given prior work on the development of children’s use of language in teaching (Strauss et al., 2002) and the role of linguistic development on false-belief understanding (Cutting & Dunn, 1999; Lohmann & Tomasello, 2003). Children’s relatively poor performance on our more language-dependent

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

novel word task provides tentative evidence that language may be a particularly important measure for evidence selection in word learning tasks. Future work could investigate whether the link between pedagogical evidence selection and ToM is in any way moderated by language development.

In Experiment 3, we trained pedagogical skill to improve children's false-belief understanding, and found tentative evidence that children with the most room for false-belief improvement benefitted from the training. This raises exciting questions about the precise relationship between these two mechanisms: How is information about selecting evidence and others' minds integrated? Are these processes shared across all domains, or tailored to particular kinds of tasks? Do these abilities develop in tandem, or might one depend on the development of the other? Our goal was to begin to answer these questions with a controlled training task, but future work could investigate the causal relationship between these domains by exploring whether training false-belief understanding also leads to improved efficacy of pedagogical evidence selection.

Here, we used false-belief understanding to measure ToM development. Of course, there are also many other aspects of ToM reasoning that we chose not to assess – for instance, while even infants make implicit inferences about others' desires (e.g., Behne, Liszkowski, Carpenter, & Tomasello, 2012; Repacholi & Gopnik, 1997), the ability to predict others' actions or emotions based on their beliefs does not emerge until much later in development (Bartsch & Wellman, 1989; Wellman & Liu, 2004). Representing another's knowledge state, and the evidence required to revise it, has clear ties to representing false-beliefs, but may not be as strongly linked to different facets of ToM, such as recognizing contradictory emotions. Similarly, we operationalized teaching skills as effective evidence selections, but this is just one

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

of many components that make up “good teaching” – which has been measured in past work using tasks that are quite different from those used here (e.g., Davis-Unger & Carlson, 2008). Future work should investigate these more comprehensive measures of ToM and teaching abilities, in order to deepen our understanding of how exactly various components of ToM reasoning relate to different elements of effective teaching throughout development.

It should be acknowledged that our sample was limited to children reared within the United States; it is therefore unclear how our findings might generalize to other cultures. In some Mayan communities, for instance, children become proficient teachers at earlier ages than they do in Western cultures (Gaskins, 1999; Maynard, 2002; see also Shneidman, Gweon, Schulz, & Woodward, 2016). If ToM development and teaching abilities are indeed linked as we report here, this could have implications for how ToM develops across cultures where natural pedagogy emerges at different ages. Important directions for future work might therefore involve expanding the populations of children participating in these kinds of studies to shed further insight on how the development of ToM abilities is tied to the age at which natural pedagogy tends to emerge.

Our results may also speak to the development of children's expectations about the role of teachers as evidence selectors in pedagogical contexts. It is of particular note that while children here spontaneously selected useful examples for their learners, there is some evidence to suggest that children's own learning is enhanced in contexts that involve learner-led discovery and exploration (e.g., Alfieri, Brooks, Aldrich, & Tenenbaum, 2011; Fisher, Hirsh-Pasek, Newcombe, & Golinkoff, 2013). Investigating how children develop expectations about when self-discovery leads to superior learning outcomes, and whether they are able to incorporate these strategies into their own teaching, would be a fascinating direction for future work. It will

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

also be critical to explore how these expectations translate to formal schooling environments in later childhood. It is possible that the current findings might even be of some use for educational professionals: Perhaps our perspective on young children's reasoning about pedagogically sampled evidence could contribute to the field's conceptualizations of how learning may take place in the classroom during these formative years.

Our goal was to explore the development of preschoolers' ability to select pedagogical evidence in the service of teaching, and also to establish a link between this skill and ToM proficiency. We found evidence for a bidirectional relationship between the ability to reason about others' beliefs and the efficacy of children's evidence selections. Our paper may also speak to broader theories of natural pedagogy – suggesting a potential link between the uniquely human ability to teach others and the development of the ability to reason about others' minds. This raises questions about whether these skills are cognitively intertwined due to similar evolutionary pressures: Perhaps the many social cooperative skills that seem to set humans apart from other primates (e.g., Boyd, Richerson, & Henrich, 2011; Moll & Tomasello, 2007; Tomasello, Carpenter, Call, Behne, & Moll, 2005) are inextricably connected throughout development. That great apes are unable to understand the notion of false beliefs under cooperative contexts (e.g., Kaminski, Call, & Tomasello, 2008) could link to their lacking spontaneous pedagogy and provides a clearer story on how the presence of these linking abilities sets humans apart. Whatever the case may be, reasoning about other minds, as conceptualized in the field, is composed of multiple interrelated inference problems. Understanding the role of these social inferences in learning requires investigating how children approach several conjoined problems. We hope that the studies demonstrated here provide a relevant example of the value of that approach.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

References

- Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology, 103*(1), 1-18.
<https://doi.org/10.1037/a0021017>
- Alvarez, A. L., & Booth, A. E. (2015). Preschoolers prefer to learn causal information. *Frontiers in Psychology, 6*, 60. <https://doi.org/10.3389/fpsyg.2015.00060>
- Amsterlaw, J., & Wellman, H. M. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development, 7*(2), 139-172. https://doi.org/10.1207/s15327647jcd0702_1
- Ashley, J., & Tomasello, M. (1998). Cooperative problem-solving and teaching in preschoolers. *Social Development, 7*(2), 143-163. <https://doi.org/10.1111/1467-9507.00059>
- Astington, J. W., & Jenkins, J. M. (1995). Theory of mind development and social understanding. *Cognition and Emotion, 9*(2-3), 151-165.
<https://doi.org/10.1080/02699939508409006>
- Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences, 14*(3), 110-118. <https://doi.org/10.1016/j.tics.2009.12.006>
- Baker, S. T., Leslie, A. M., Gallistel, C. R., & Hood, B. M. (2016). Bayesian change-point analysis reveals developmental change in a classic theory of mind task. *Cognitive Psychology, 91*, 124-149. <https://doi.org/10.1016/j.cogpsych.2016.08.001>
- Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition, 118*(1), 84-93.
<https://doi.org/10.1016/j.cognition.2010.10.010>

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Bartsch, K., & Wellman, H. (1989). Young children's attribution of action to beliefs and desires.

Child Development, 60(4), 946-964. <https://doi.org/10.2307/1131035>

Behne, T., Liszkowski, U., Carpenter, M., & Tomasello, M. (2012). Twelve-month-olds'

comprehension and production of pointing. *British Journal of Developmental*

Psychology, 30(3), 359-375. <https://doi.org/10.1111/j.2044-835X.2011.02043.x>

Bonawitz, E. B., Fischer, A., & Schulz, L. (2012). Teaching 3.5-year-olds to revise their beliefs

given ambiguous evidence. *Journal of Cognition and Development*, 13(2), 266-280.

<https://doi.org/10.1080/15248372.2011.577701>

Booth, A. E. (2009). Causal supports for early word learning. *Child Development*, 80(4), 1243-

1250. <https://doi.org/10.1111/j.1467-8624.2009.01328.x>

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is

essential for human adaptation. *Proceedings of the National Academy of Sciences*,

108(Supplement 2), 10918-10925. <https://doi.org/10.1073/pnas.1100290108>

Cutting, A. L., & Dunn, J. (1999). Theory of mind, emotion understanding, language, and family

background: Individual differences and interrelations. *Child Development*, 70(4), 853-

865. <https://doi.org/10.1111/1467-8624.00061>

Davis-Unger, A. C., & Carlson, S. M. (2008). Development of teaching skills and relations to

theory of mind in preschoolers. *Journal of Cognition and Development*, 9(1), 26-45.

<https://doi.org/10.1080/15248370701836584>

Eaves, B., & Shafto, P. (2012). Unifying Pedagogical Reasoning and Epistemic Trust. In F. Xu

& T. Kushnir (Eds.), *Advances in child development and behavior* (pp. 295-319). San

Diego, CA: Elsevier. <https://doi.org/10.1016/b978-0-12-397919-3.00011-3>

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Eaves, B., & Shafto, P. (2017). Parameterizing developmental changes in epistemic trust.

Psychonomic Bulletin & Review, 24(2), 277-306. <https://doi.org/10.3758/s13423-016-1082-x>

Fisher, K. R., Hirsh-Pasek, K., Newcombe, N., & Golinkoff, R. M. (2013). Taking shape:

Supporting preschoolers' acquisition of geometric knowledge through guided play. *Child Development*, 84(6), 1872-1878. <https://doi.org/10.1111/cdev.1209>

Gaskins, S. (1999). Children's daily lives in a Mayan village: A case study of culturally

constructed roles and activities. In A. Goncu (Ed.), *Children's engagement in the world* (pp 25-81). Cambridge, England: Cambridge University Press.

<https://doi.org/10.1177/106939710003400405>

Gleason, J. B., & Ratner, N. B. (2009). *The development of language*. Boston, MA: Pearson.

Goodman, N. D., Baker, C. L., Bonawitz, E., Mansinghka, V. K., Gopnik, A., Wellman, H.,

Schulz, L., & Tenenbaum, J. (2006). Intuitive theories of mind: A rational approach to false belief. *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*.

Guasti, M. T. (2002). *Language acquisition: The growth of grammar*. Cambridge, MA: MIT Press.

Gweon, H., Tenenbaum, J., & Schulz, L. (2010). Infants consider both the sample and the

sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, 107(20), 9066-9071. <https://doi.org/10.1073/pnas.1003095107>

Holyoak, K. J., Junn, E. N., & Billman, D. O. (1984). Development of analogical problem-

solving skill. *Child Development*, 55(6), 2042-2055. <https://doi.org/10.2307/1129778>

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

- Imuta, K., Henry, J. D., Slaughter, V., Selcuk, B., & Ruffman, T. (2016). Theory of mind and prosocial behavior in childhood: A meta-analytic review. *Developmental Psychology*, 52(8), 1192–1205. <https://doi.org/10.1037/dev0000140>
- Kalish, C. W., & Lawson, C. A. (2007). Negative evidence and inductive generalisation. *Thinking & Reasoning*, 13(4), 394-425. <https://doi.org/10.1080/13546780701273402>
- Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, 109(2), 224-234. <https://doi.org/10.1016/j.cognition.2008.08.010>
- Knudsen, B., & Liszkowski, U. (2012). One-year-olds warn others about negative action outcomes. *Journal of Cognition and Development*, 14(3), 424-436. <https://doi.org/10.1080/15248372.2012.689387>
- Koerber, S., Sodian, B., Thoermer, C., & Nett, U. (2005). Scientific reasoning in young children: Preschoolers' ability to evaluate covariation evidence. *Swiss Journal of Psychology*, 64(3), 141-152. <https://doi.org/10.1024/1421-0185.64.3.141>
- Kushnir, T., & Koenig, M. (2017). What I don't know won't hurt you: The relation between professed ignorance and later knowledge claims. *Developmental Psychology*, 53(5), 826-835. <https://doi.org/10.1037/dev0000294>
- Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*, 81(3), 929-944. <https://doi.org/10.1111/j.1467-8624.2010.01443.x>
- Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development*, 74(4), 1130-1144. <https://doi.org/10.1111/1467-8624.00597>

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

- Maynard, A. E. (2002). Cultural teaching: The development of teaching skills in Maya sibling interactions. *Child Development*, 73(3), 969-982. <https://doi.org/10.1111/1467-8624.00450>
- 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>
- Moll, H., & Tomasello, M. (2007). Cooperation and human cognition: The Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1480), 639-648. <https://doi.org/10.1098/rstb.2006.2000>
- Nurmsoo, E., & Robinson, E. J. (2009). Children's trust in previously inaccurate informants who were well or poorly informed: When past errors can be excused. *Child Development*, 80(1), 23-27. <https://doi.org/10.1111/j.1467-8624.2008.01243.x>
- Olson, D. R., & Bruner, J. S. (1996). Folk psychology and folk pedagogy. In D. R. Olson & N. Torrance (Eds.), *The handbook of education and human development: New models of learning, teaching and schooling* (pp. 9-27). Malden, MA: Blackwell Publishing. <https://doi.org/10.1111/b.9780631211860.1998.00003.x>
- Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics-pragmatics interface. *Cognition*, 86(3), 253-282. [https://doi.org/10.1016/S0010-0277\(02\)00179-8](https://doi.org/10.1016/S0010-0277(02)00179-8)
- Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, 43(5), 1216-1226. <https://doi.org/10.1037/0012-1649.43.5.1216>

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Piaget, J. (1952). *The child's conception of number*. New York, NY: Norton.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental Psychology*, 33(1), 12. <https://doi.org/10.1037/0012-1649.33.1.12>

Rhodes, M., Bonawitz, E., Shafto, P., Chen, A., & Caglar, L. (2015). Controlling the message: Preschoolers' use of information to teach and deceive others. *Frontiers in Psychology*, 6, 867. <https://doi.org/10.3389/fpsyg.2015.00867>

Rhodes, M., Gelman, S. A., & Brickman, D. (2010). Children's attention to sample composition in learning, teaching and discovery. *Developmental Science*, 13(3), 421- 429. <https://doi.org/10.1111/j.1467-7687.2009.00896.x>

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science*, 15(3), 436-447. <https://doi.org/10.1111/j.1467-7687.2012.01135.x>

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71, 55-89. <https://doi.org/10.1016/j.cogpsych.2013.12.004>

Shafto, P. & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*.

Shneidman, L., Gweon, H., Schulz, L. E., & Woodward, A. L. (2016). Learning from others and spontaneous exploration: A cross-cultural investigation. *Child Development*, 87(3), 723-735. <https://doi.org/10.1111/cdev.12502>

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

- Slaughter, V., & Gopnik, A. (1996). Conceptual coherence in the child's theory of mind: Training children to understand belief. *Child Development, 67*(6), 2967-2988.
<https://doi.org/10.1111/j.1467-8624.1996.tb01898.x>
- Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Language Learning and Development, 11*(2), 176-190.
<https://doi.org/10.1080/15475441.2014.927328>
- Strauss, S., Ziv, M., & Stein, A. (2002). Teaching as a natural cognition and its relations to preschoolers' developing theory of mind. *Cognitive Development, 17*(3), 1473-1487.
[https://doi.org/10.1016/S0885-2014\(02\)00128-4](https://doi.org/10.1016/S0885-2014(02)00128-4)
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*(4), 629-640.
<https://doi.org/10.1017/S0140525X01000061>
- Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 721-727. <https://doi.org/10.1017/S0140525X05000129>
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*(3), 655-684.
<https://doi.org/10.1111/1467-8624.00304>
- Wellman, H. M., & Lagattuta, K. H. (2004). Theory of mind for learning and teaching: The nature and role of explanation. *Cognitive Development, 19*(4), 479-497.
<https://doi.org/10.1016/j.cogdev.2004.09.003>
- Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75*(2), 523-541. <https://doi.org/10.1111/j.1467-8624.2004.00691.x>

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*(2), 343-357.

<https://doi.org/10.1016/j.cognition.2014.07.008>

Wood, D., Wood, H., Ainsworth, S., & O'Malley, C. (1995). On becoming a tutor: Toward an ontogenetic model. *Cognition and Instruction*, *13*(4), 565-581.

http://doi.org/10.1207/s1532690xci1304_7

Ziv, M., & Frye, D. (2004). Children's understanding of teaching: The role of knowledge and belief. *Cognitive Development*, *19*(4), 457-477.

<https://doi.org/10.1016/j.cogdev.2004.09.002>

Ziv, M., Solomon, A., Strauss, S., & Frye, D. (2016). Relations between the development of teaching and theory of mind in early childhood. *Journal of Cognition and Development*, *17*(2), 264-284. <https://doi.org/10.1080/15248372.2015.1048862>

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

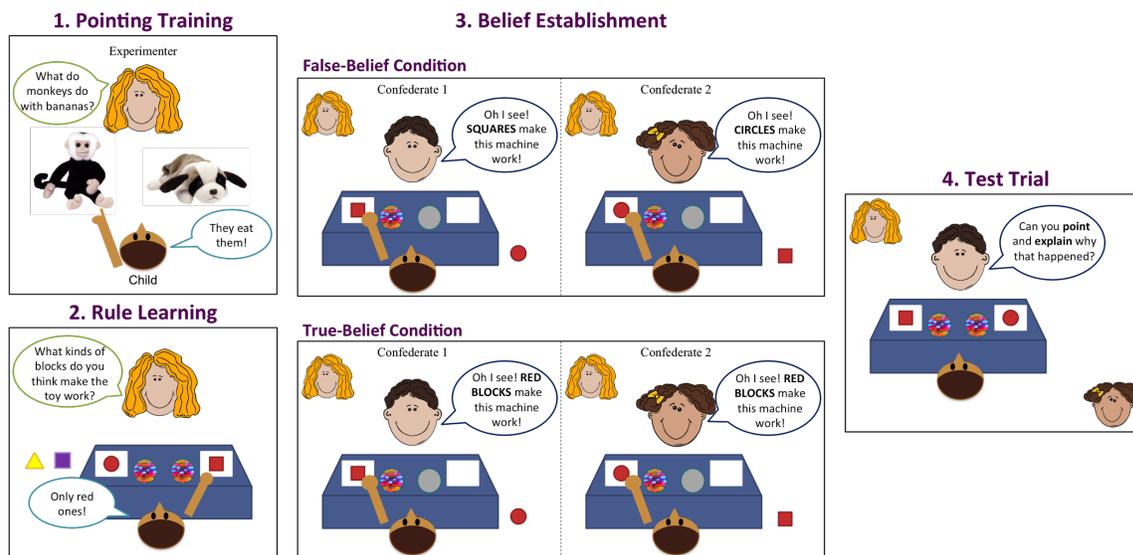


Figure 1. A schematic of the procedure used for Experiment 1.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

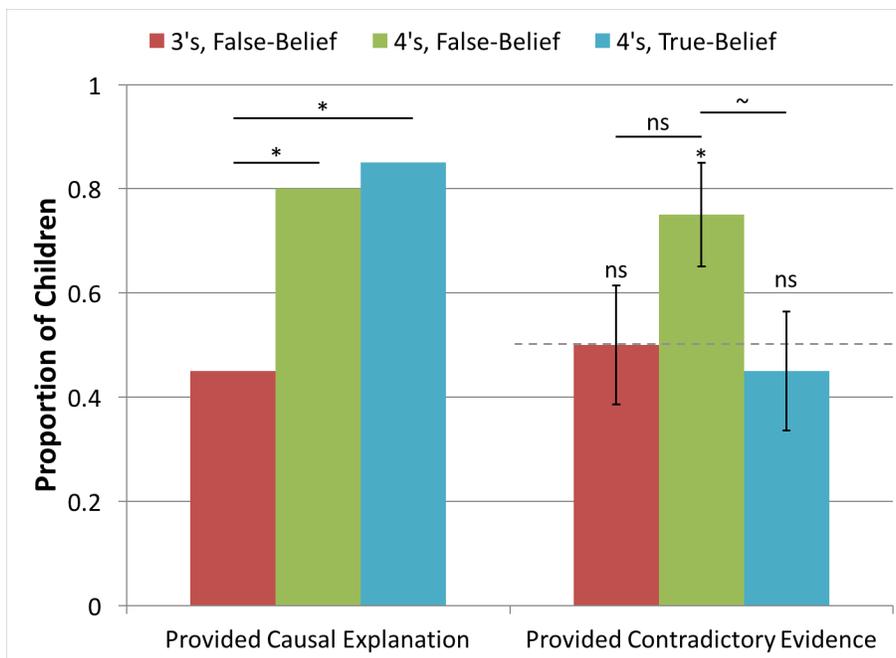


Figure 2. The proportion of three-year-olds and four-year-olds in the False-Belief and True-Belief conditions who: 1) provided a causal explanation when pointing; and 2) pointed to the block that the confederate had not seen in Experiment 1. Asterisks denote significant differences at the $p < .05$ level; the tilde represents significance at the $p < .10$ level (two-tailed). Error bars represent (+/-) the standard error. Chance is represented by the dashed line (50%).

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

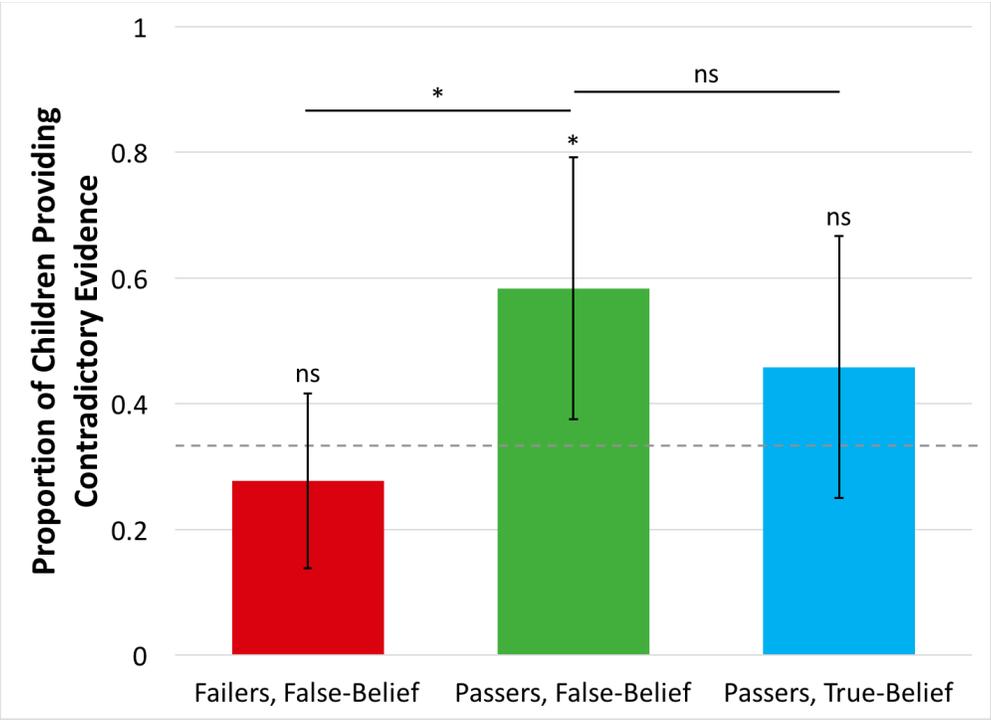


Figure 3. The proportion of false-belief failers and passers in the False-Belief and True-Belief conditions who selected the block that the confederate had not seen on the first test trial in Experiment 2. Error bars represent bootstrapped 95% confidence intervals. Chance is represented by the dashed line (33%).

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION



Figure 4. A schematic of our study design for Experiment 3, with examples of possible pedagogical training schedules for false-belief passers and failers.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

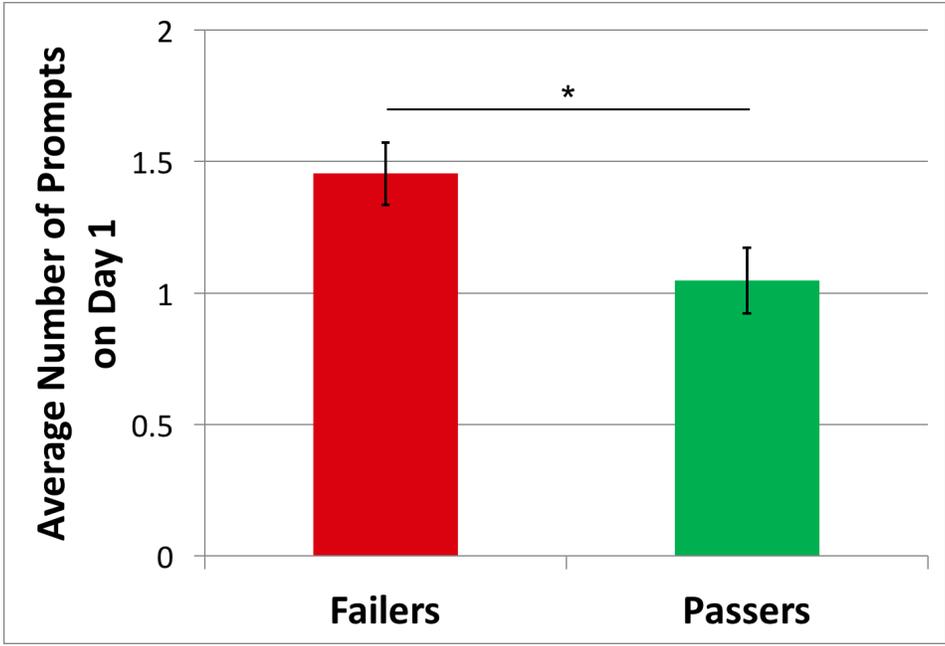


Figure 5. The average number of prompts required across both pedagogical tasks on the preliminary testing day for children who passed vs. failed the false-belief battery at pre-test in Experiment 3 (lower scores indicate better performance). Error bars represent (+/-) the standard error.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

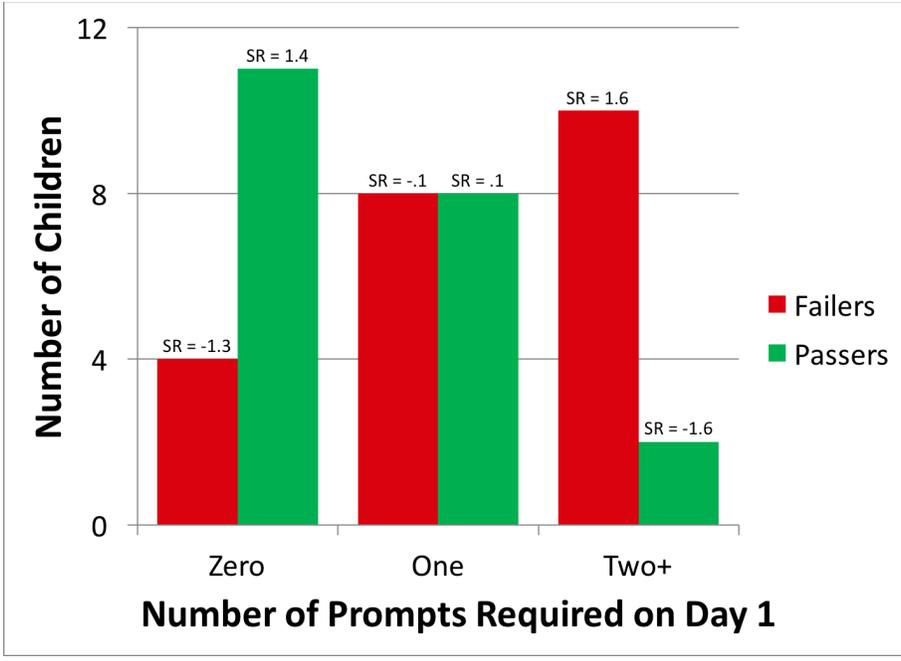


Figure 6. The number of false-belief passers and failers who required zero, one, and two or more prompts during the causal toy task on the preliminary testing day in Experiment 3. Standardized residuals (SR) for each of these frequencies are depicted.

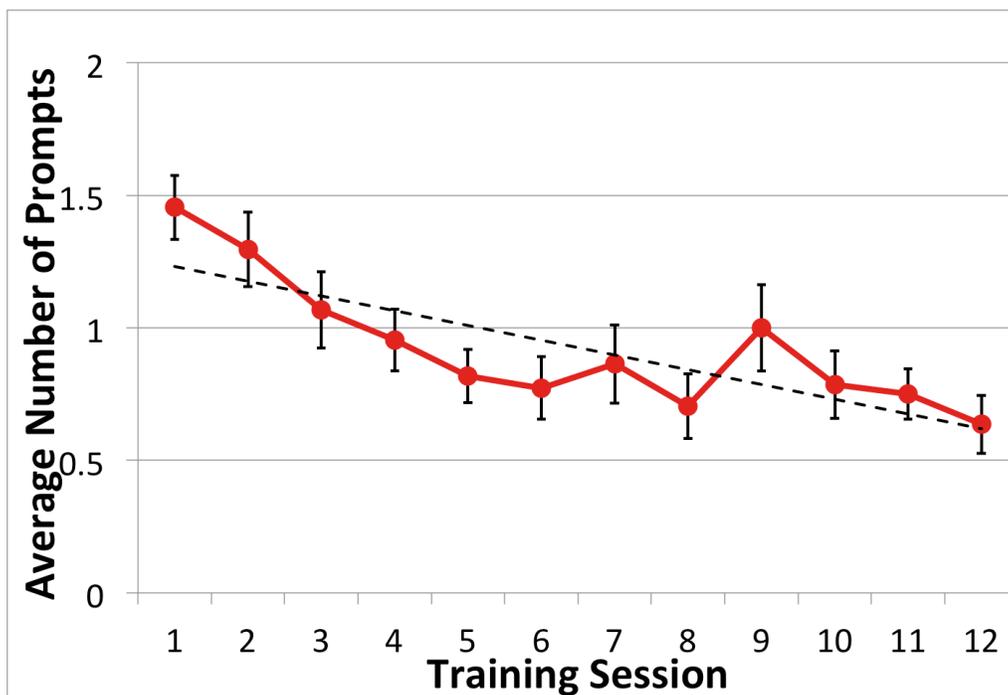


Figure 7. The average number of prompts required by failers in the training condition on each session in Experiment 3. Error bars represent (+/-) the standard error.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

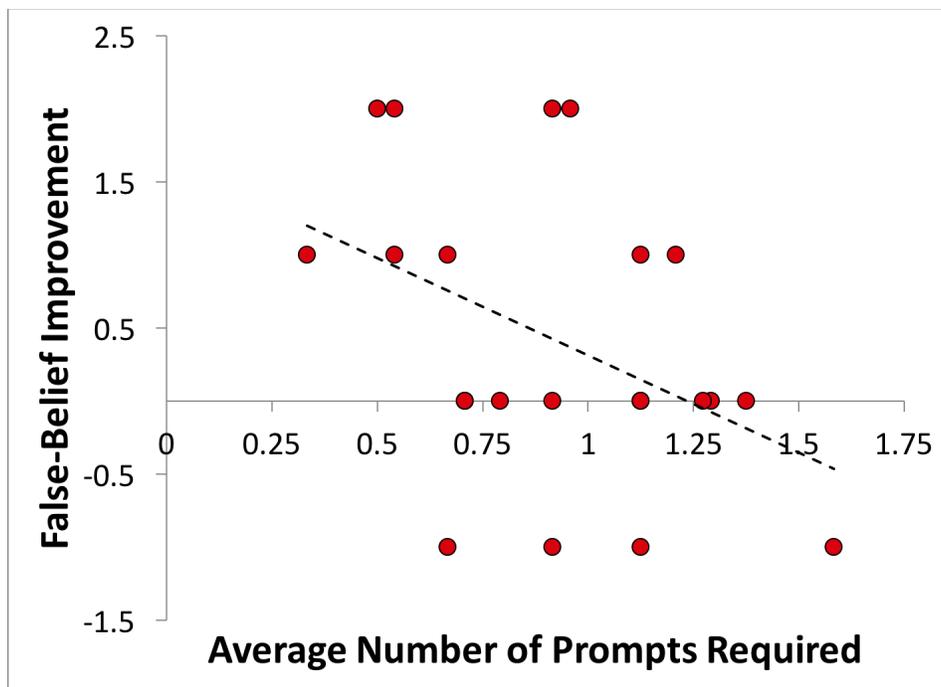


Figure 8. The average false-belief improvement of failers in the training condition, correlated with the average number of prompts required over the course of the pedagogical training in Experiment 3.

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

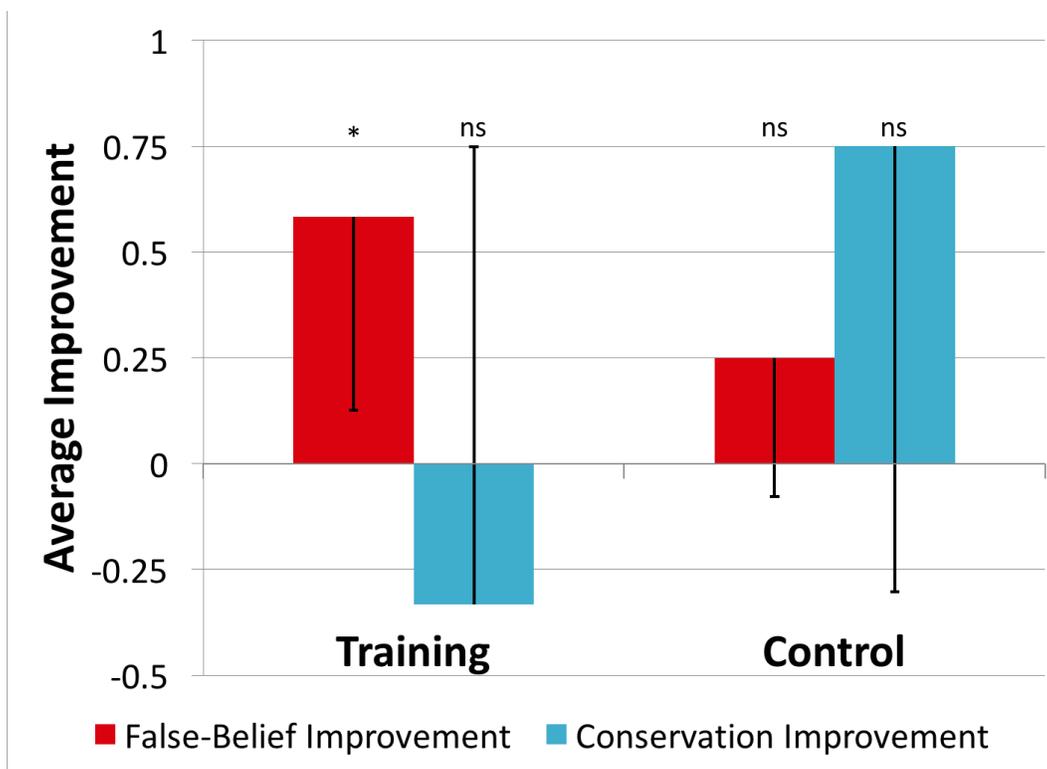


Figure 9. Average false-belief and conservation improvement for children who answered all false-belief questions incorrectly at pre-test in Experiment 3. Error bars represent two times the standard error.

CHILDREN’S THEORY OF MIND AND EVIDENCE SELECTION

SUPPLEMENTAL MATERIALS

Experiment 1

The number of 3-year-olds and 4-year-olds across both experimental conditions in Experiment 1 who provided causal explanations, descriptive explanations, or failed to provide an explanation for why the toy activated.

	Causal	Descriptive	Failure to Respond
3's: False-Belief	9	6	5
4's: False-Belief	16	3	1
4's: True-Belief	17	2	1

Experiment 2

This experiment is preregistered with the Open Science Framework at the following link:

<https://osf.io/4fztv/register/565fb3678c5e4a66b5582f67>

The causal toy and blocks used in Experiment 2:



CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Experiment 3**Pre- and Post-test assessments**

Task	Stimuli
Numerical conservation, Trial 1	Pre-test: 10 pennies arranged in two rows Post-test: 10 washers arranged in two rows
Numerical conservation, Trial 2	Pre-test: 10 polyhedrons arranged in two rows Post-test: 10 blocks arranged in two rows
False Belief, Sally-Anne	Pre-test: Story Book 1 (Sally hides teddy bear in basket; Alex moves to box) Post-test: Story Book 2 (Billy hides cookie in drawer; Anne moves to cabinet)
False Belief, Unexpected contents	Pre-Test: Band-Aid box with pencil in it Post-Test: Crayon box with keys in it

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Causal Toy Task

Toy Name	Toy Mechanisms	Activation Rule	Correct Response
Red Airplane	Wheel/bell	Anytime the bell is rung	<ol style="list-style-type: none"> 1. Show that toy activates when bell is rung 2. Show that toy does not activate when wheel is spun
Shadowbox	Dropping chips into basket/cup	Anytime there's a chip in the cup	<ol style="list-style-type: none"> 1. Show that toy activates when chip is in the cup 2. Show that toy does not activate when chip is in the basket
Phone	Red chip/blue chip on an activator pad	Anytime the blue chip is on	<ol style="list-style-type: none"> 1. Show that toy activates when blue chip is on 2. Show that toy does not activate when red chip is on
Helicopter	Sliding black block/white block	Anytime the black one slides	<ol style="list-style-type: none"> 1. Show that toy activates when black one slides 2. Show that toy does not activate when white one slides
Gear Toy	Putting red wheel/blue wheel on toy	Need to put both the red wheel and the blue wheel on	<ol style="list-style-type: none"> 1. Show that toy activates when both red wheel and blue wheel are on 2. Show that toy does not activate when just red wheel is on 3. Show that toy does not activate when just blue wheel is on
Purple	Spinning blue wheel/pink wheel	Need to spin both blue wheel and pink wheel at the same time	<ol style="list-style-type: none"> 1. Show that toy activates when both blue wheel and pink wheel are spun 2. Show that toy does not activate when just blue wheel is spun 3. Show that toy does not activate when just pink wheel is spun

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Novel Word Task

Novel Word	Picture Features	Word Meaning	Correct Response
Wug	Blue car*, red car, blue truck, red truck**	Car	<ol style="list-style-type: none"> 1. Show that blue car is Wug 2. Show that red car is Wug 3. Show that blue truck is not Wug
Blicket	Black cat*, white cat, black dog, white dog**	Black	<ol style="list-style-type: none"> 1. Show that black cat is Blicket 2. Show that black dog is Blicket 3. Show that white cat is not Blicket
Fep	Striped shirt*, striped pants, solid shirt, solid pants**	Striped	<ol style="list-style-type: none"> 1. Show that striped shirt is Fep 2. Show that striped pants is Fep 3. Show that solid shirt is not Fep
Modi	Fuzzy triangle*, fuzzy square, plain triangle, plain square**	Triangle	<ol style="list-style-type: none"> 1. Show that fuzzy triangle is Modi 2. Show that plain triangle is Modi 3. Show that fuzzy square is not Modi
Dax	Black spoon*, black fork, white spoon, white fork**	Black spoon	<ol style="list-style-type: none"> 1. Show that black spoon is Dax 2. Show that white spoon is not Dax 3. Show that black fork is not Dax
Toma	Red chip*, blue chip, red fabric, blue fabric**	Red chip	<ol style="list-style-type: none"> 1. Show that red chip is Toma 2. Show that blue chip is not Toma 3. Show that red fabric is not Toma

*Original picture presented to represent novel word

**Distractor item that was removed before test trial

CHILDREN'S THEORY OF MIND AND EVIDENCE SELECTION

Examples of stimuli used in causal toy task:

Red Airplane

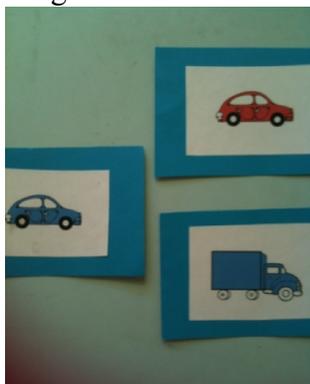


Shadowbox



Examples of stimuli used in novel word task:

Wug



Blicket

